Advanced Physical Activity Classification Using Wearable Sensor Data and Machine Learning

Table of Contents

1. Introduction

1.1. Project Overview

The Colibri Wireless unit, represents a significant advancement in inertial measurement technology. It consists of three integrated sensors to measure acceleration, angular rate, and magnetic field, along with separate sensors for temperature and orientation. In this project, focus is to harness the potential of the Colibri unit to classify different physical activities accurately.

1.2. Research Objective

The primary objective of this research is to develop a robust predictive model using machine learning techniques. This model will utilize the data gathered from the IMUs and heart-rate monitors to distinguish between various physical activities such as walking, cycling, and playing football. Successful classification and prediction of these activities will not only demonstrate the effectiveness of the Colibri unit but also provide actionable insights for its optimization in real-world health and fitness applications.

1.3. Data Description

The study involves data collected from nine subjects who each participated in 12 core physical activities, with some conducting an additional six. The data, stored in individual .dat files, contains 54 attributes including timestamps, activity IDs, heart rate measurements, and IMU readings from sensors attached to the wrist, chest, and ankle.

2. Contextual Background

2.1. Wearable Technology in Physical Activity Monitoring

The rise of wearable technology has revolutionized the way we monitor and understand physical activity and health. Devices like the Trivisio Colibri Wireless unit are at the forefront of this evolution, offering detailed insights into human movement and physiological metrics. These devices have applications ranging from personal fitness tracking to professional sports analytics and healthcare monitoring.

2.2. The Importance of Accurate Activity Classification

Accurately classifying physical activities is crucial in numerous sectors. For individuals, it helps in tracking and improving fitness levels and in monitoring health parameters. In professional sports, precise activity classification can aid in performance optimization and injury prevention. In healthcare, it assists in patient monitoring, rehabilitation, and research into physical activity's impact on various health conditions.

2.3. The Challenge

Despite the advancements in technology, accurately distinguishing between different types of physical activities using sensor data remains a complex challenge. Factors like sensor placement, individual differences in movement, and the subtleties of different activities make this a non-trivial task. The Trivisio Colibri Wireless unit, with its advanced sensor array, offers a potential solution to these challenges.

2.4. Research Implications

This study is not just about proving the technical prowess of the Colibri unit. It's about pushing the boundaries of what's possible in wearable technology for physical activity monitoring. The insights gained from this research have the potential to influence future designs and functionalities of wearable devices, making them more intuitive, accurate, and user-friendly.

3. Hypothesis

It is hypothesized that by examining the changes in heart rate and specific readings from motion-tracking sensors (IMUs), different physical activities can be accurately identified. It is assumed that unique patterns are exhibited by each physical activity, such as walking, running, or cycling. These patterns, characterized by variations in heart rate and specific movements, are captured by motion sensors (IMUs) attached to the hand, chest, and ankle. The aim is to analyze these distinct signatures and develop a predictive model. This model will not only differentiate between various activities but also demonstrate the effectiveness of the Trisivio Colibri Wireless unit in capturing these nuances, as compared to its competitors.

The aim is to prove the effectiveness of the Trivisio Colibri Wireless unit. The successful identification of activities using data from this unit would demonstrate its ability to accurately capture key health and movement information.

Information gathered from IMUs is the starting point. These advanced devices are capable of recording human movement in fine detail. Readings are encompassed from three distinct points of wear: the hand, chest, and ankle, providing a comprehensive view of physical activity. The dataset is structured to include 'time', 'activityID', and 'heart rate', along with numerous sensor readings from the IMUs, named 'IMU_hand_x', 'IMU_chest_x', and 'IMU_ankle_x', where 'x' ranges from 1 to 17.

The layout of the data is initially set up to match the detailed structure of the information gathered. Data is loaded and prepared from a collection of .dat files, each representing a piece of a larger puzzle in this study. These parts are combined into a single large data frame, with careful attention paid to maintaining the correct order and quality of the sensor records.

In navigating the preprocessing phase, a decision is made to discard transient activities, identified by 'activityID=0', as they are not contributory to the core objective of the analysis. Attention is then directed towards addressing missing values within the dataset. Given the critical importance of complete data for accurate modeling, missing values are imputed with the mean of their respective columns, thus ensuring a robust foundation for the subsequent analytical processes.

4. Data Overview

The Trivisio Colibri Wireless unit's performance is assessed through real-world activity data, drawn from:

- **Protocol Activities:** Standard exercises, providing baseline performance data.

- **Optional Activities:** Extra exercises to broaden the dataset for deeper analysis.

Data, recorded in .dat files, represents individual activity sessions. These files create an extensive dataset that includes timestamps, activity classification, heart rate in bpm, and IMU readings from the hand, chest, and ankle. These details are crucial for activity recognition and analysis. IMU data is especially vital as it showcases the unit's ability to track movement, a key differentiator from competitors.

4.1. Data Cleaning Justification

The data cleaning process undertaken in this project is guided by both domain knowledge and specific assumptions to ensure the reliability and relevance of the data for our analysis.

Outlier Removal for Heart Rate Data: Based on physiological insights, heart rate readings that fall outside the typical range of 40 bpm (resting) to 200 bpm (peak exercise) are identified as outliers. These values are considered physiologically implausible for the average individual. The removal of these outliers is crucial to maintain the integrity of the dataset, as they can skew the results and impair the model's ability to make accurate predictions.

Standardization of Numerical Columns: The varied scales of sensor readings pose a challenge for comparative analysis. To address this, the StandardScaler is employed to normalize the data. This standardization process, which scales data to have a mean of 0 and a standard deviation of 1, reduces the model's sensitivity to the scale of features. This step is essential for enhancing the model's ability to learn patterns from the data effectively.

Introduction of Heart Rate Variability Feature: The decision to engineer a feature that captures the variability in heart rate is based on the understanding that different physical activities induce varying levels of fluctuation in heart rate. This variability is a potentially valuable predictor for differentiating between types of activities, as certain activities might induce more significant fluctuations compared to others.

These data cleaning steps are integral to preparing the dataset for subsequent analysis. By ensuring that the data is clean, standardized, and enhanced with relevant features, the foundation is laid for developing a robust and accurate predictive model.

4.2. Rationale For Data Cleaning

Outlier Removal in Heart Rate Data:

Heart rate readings that fall outside the range of 40 bpm (typical resting heart rate) to 200 bpm (peak exercise level) are generally considered physiologically implausible for most individuals. The removal of these outliers is undertaken to ensure the dataset reflects realistic heart rate values. This enhances the model's capacity for generalization and accurate predictions in likely scenarios. Records where the heart rate falls outside this defined range are identified and excluded. Such practice is standard in data preprocessing, particularly with biological data, to mitigate the impact of erroneous readings on the analysis.

Standardization of Numerical Columns:

Sensors operating on varying scales are observed, and the standardization of these readings is carried out to allow a more uniform comparison across different measurements. Standardization, involving scaling data to have a mean of 0 and a standard deviation of 1, is applied to reduce the model's sensitivity to the scale of features, thereby enhancing its pattern learning effectiveness. The StandardScaler is applied to numerical columns to normalize their distributions, with means and standard deviations preserved for future uses, such as model deployment on new data, to ensure consistency in data handling.

Feature Engineering - Heart Rate Variation:

Variability in heart rate over time is recognized as an important factor in distinguishing between different types of physical activities. Activities that induce more fluctuation in heart rate than others are particularly noted. The introduction of a new feature, capturing the standard deviation of heart rate over a rolling window, is undertaken. This addition potentially enhances the dataset with a valuable predictor, likely to improve the model's accuracy.

5. Final EDA Summary

The exploratory data analysis undertaken herein meticulously examines the heart rate distributions and sensor readings to uncover the physiological and movement patterns corresponding to various physical activities. In the dataset, patterns within physical activity data are uncovered through statistical visualizations. Heart rate distributions for each activity are displayed via histograms, essential for identifying differential responses to exercises. Relationships between IMU sensor readings are revealed through correlation studies, highlighting connections and distinct data contributions. Sensor data is contrasted with heart rate variations using bivariate pairplots and boxplots, which elucidate trends and the diversity of movements across activities. Finally, a correlation matrix, accompanied by p-values, quantifies these relationships, providing a foundational backdrop for the ensuing predictive modeling.

5.1. Heart Rate Distributions

The heart rate distribution graphs corroborate the initial interpretation, presenting distinct peaks that align with typical physiological responses to different activities. The narrow distribution for some activities indicates a uniform exertion level, while the broader distributions suggest variability in intensity. This variability is paramount for identifying activities with fluctuating exertion levels. The heart rate data, standardized with a mean of zero and a standard deviation of one, largely falls within the expected range, validating the process of outlier management and indicating the data's readiness for predictive modeling.

5.2. Correlation Matrix of IMU Readings

The correlation heatmap reinforces the initial interpretation by revealing significant relationships between sensor readings, which could reflect synchronized movements across body segments. The visualization of low correlation coefficients between certain sensor pairings underscores the diverse information captured by the sensors, enhancing the potential for these readings to serve as distinctive features in activity classification models. This diversity in sensor data is crucial for a model's ability to differentiate between activities with similar motion profiles.

5.3. Pairplot Relationships and Sensor Data Insights

The pairplot further substantiates the relationships highlighted in the initial interpretation by providing a bivariate view that brings to light the clustering of activities and the inter-variable relationships. The scatter plots and histograms from the pairplot offer additional granularity, revealing how different activities cluster together and how heart rate variation interacts with sensor data.

5.4. Boxplot Comparisons

Boxplots across activities for key sensor readings offer a visual comparison of the medians, spreads, and outliers, contributing to a deeper understanding of the physical movements characteristic of each activity. These plots are essential for recognizing the range and variability of movements as captured by the IMUs.

Integrating these observations, the EDA affirms that the Trivisio Colibri Wireless unit's sensor readings, along with heart rate data, provide a comprehensive dataset capable of distinguishing between various physical activities. The standardized heart rate and the diverse sensor readings lay a strong foundation for predictive modeling. The correlation matrix, with its detailed insights into the relationships between sensors, will be instrumental in feature selection, enhancing the model's accuracy and interpretability.

The robustness and comprehensiveness of the predictive model will benefit from these exploratory insights, as they ensure that the features selected for the model encapsulate the essential patterns and nuances present in the physical activity data.

6. Feature Engineering

Defining Feature Columns: In combined_df, columns to be used as features are identified, excluding 'activityID', which is the target variable. This ensures that focus is placed only on essential columns for further analysis.

Handling Missing Values with Imputation: Missing values within the feature columns are addressed using the SimpleImputer, which replaces missing values with the average of each column (strategy='mean'). This approach is commonly employed for numerical data to maintain data integrity, ensuring that no values are discarded due to their absence.

Encoding the Categorical Target Variable: The target variable, 'activityID', is transformed using LabelEncoder. This conversion is crucial for translating categorical data into numerical format, suitable for processing by machine learning algorithms.

Feature Normalization: MinMaxScaler is utilized to adjust feature values to a range between 0 and 1. This normalization is important across various machine learning models as it ensures that all features contribute equally to the model training, regardless of their original scale.

Feature Selection Using SelectKBest: SelectKBest, employing the chi2 score, is used for selecting the top k features. The chi-squared test, a statistical measure, is applied to determine the independence of two events. This is particularly useful in identifying features that have the strongest association with the target variable. A total of 10 features (k) are selected, which can be adjusted based on data and model requirements.

Evaluating and Selecting Features: Following the application of SelectKBest, a data frame named 'feature_scores' is created, displaying scores for all features based on the chi-squared test. Features are then ranked by their score, and the top k features are chosen for model training. This step is integral to the model's efficacy, ensuring that only the most influential features are utilized, potentially enhancing accuracy and model robustness.

Preparation for Modeling: The selected features (X_selected) and the transformed target variable (y_encoded) are prepared for subsequent modeling stages. This preparation is pivotal in ensuring that the model is trained on well-balanced and properly processed data.

Ensuring Data Integrity Through Preprocessing: Preprocessing strategy is designed to maintain the integrity of the data. By imputing missing values, preserved valuable information that would be lost by discarding incomplete rows. The MinMaxScaler ensures that all features influence the model proportionately, preventing any single feature from disproportionately impacting the model's decisions due to scale differences.

Justification for Feature Selection: The chi-squared test serves as a statistical backbone for feature selection, emphasizing the importance of each feature's relationship with the target variable. Selection process is not only informed by statistical methods but also by domain knowledge, ensuring to capture the essence of physical activity through sensor data. The inclusion of heart rate variability, for instance, stems from its recognized importance in reflecting physical exertion, which is likely to vary significantly across different activities.

6.1. Rationale of Selected Features

The features included in the predictive model were selected based on a balanced approach, incorporating statistical significance, domain relevance, and the goal of ensuring model interpretability. These features are designed to accurately differentiate between various types of physical activities, utilizing sensor data and heart rate variability. This is particularly crucial in demonstrating the efficiency of the Trivisio Colibri Wireless unit.

6.2. Statistical Significance and Relevance

To pick the most informative features for analysis, used a method called SelectKBest, which works alongside a chi-squared test. This approach helps identify features that are crucial for predicting the type of physical activity. This statistical method identifies features that significantly correlate with the target variable, 'activityID', ensuring that the selection is grounded in statistical evidence, rather than being arbitrary. Selected features predominantly include IMU sensor readings from various body positions (hand, chest, ankle), and heart_rate_variation. The inclusion of heart_rate_variation is based on the understanding that variability in heart rate reflects different levels of physical exertion, a key characteristic for distinguishing distinct activities.

6.3. Balancing Model Performance and Interpretability

Used a technique called SMOTE to ensure our model is fair and robust, especially for activities that don't occur as often. This helps balance our data. Enabling effective performance across all types of activities, not limited to those most commonly observed. The exclusion of certain features, likely due to their low statistical significance or minimal contribution in differentiating activity types, contributes to maintaining the model's simplicity and interpretability. This selection process ensures the model's manageability and relevance, making it a reliable tool for showcasing the capabilities of the Trivisio Colibri Wireless unit in real-world applications.

7. Model Development

Employed XGBoost, a leading machine-learning algorithm known for its superior performance in classification tasks. This algorithm was pivotal in crafting a model to predict physical activities based on readings from inertial measurement units (IMUs) attached to various body parts.

Initially, the 'activityID' labels were encoded numerically to facilitate the modeling process. Key features were selected, including 'IMU_ankle_14', 'IMU_chest_14', 'IMU_chest_1', 'IMU_hand_14', 'IMU_ankle_15', 'IMU_ankle_1', 'IMU_ankle_17', 'IMU_hand_1', 'IMU_chest_16', and 'heart_rate_variation', identified through a rigorous feature selection process. These features were deemed most predictive, representing the most influential data points for training our model.

To address the class imbalance prevalent in activity recognition tasks, utilized the Synthetic Minority Over-sampling Technique (SMOTE), which artificially balances the dataset by generating synthetic samples. This preprocessing step was crucial for establishing a fair training environment for the model. Subsequently, the data was split into training and test sets to ensure a robust evaluation framework. And then leveraged GridSearchCV, a powerful tool for hyperparameter optimization, meticulously searching for the optimal combination of 'n_estimators', 'max_depth', and 'learning_rate' for the XGBClassifier, enhancing model precision and predictive capabilities.

Upon selecting the best model from GridSearchCV, we delved into its performance using classification reports. This analysis provided insights into the model's precision, recall, and

F1-scores across various activity types. Furthermore, we extracted and ranked feature importances, shedding light on how each contributed to the model's decision-making process and uncovering hidden patterns and relationships within the sensor data.

Thus, by employing XGBoost with meticulous data preparation, thoughtful feature selection, and class balancing, achieved a robust and interpretable model. The model excels not only in accurately predicting physical activities but also in offering valuable insights into the impact of sensor readings on the results.

8. Evalution of Model

The development of the model using XGBoost has demonstrated promising results. The process involved resampling the data using SMOTE to address class imbalance, ensuring that our model did not bias towards more frequent labels. Following this, we applied hyperparameter tuning via GridSearchCV, optimizing for accuracy across three folds. The best parameters identified were a learning rate of 0.1, max depth of 5, and 10 estimators.

The classification report reveals a high level of precision, recall, and f1-score across all classes, which indicates the model's robustness. Specifically, classes with a perfect score of 1.00 in precision and recall, such as classes 7, 9, and 16, demonstrate the model's capability in identifying these activities with high confidence. However, the lower scores in classes like 10 and 11 suggest there might be a need for further model refinement or additional feature engineering for these specific activities.

9. Actionable Insights

Optimizing Class-Specific Performance: The classification report about classifying things shows that the model works well for some activities but not for others. It could do better in getting the right answers and finding all the cases. For example, activities like Classes 10, 11, and 12 might have more difficult or less easy moves that the current set of traits doesn't show clearly. It would be smart to learn more about these activities. Maybe we should talk to people who know a lot about them or look at how the data is collected. Understanding the differences of these actions might show that more data from sensors is needed or it could help create special features that get a better sense of these moves.

Enhancing Feature Selection: The SHAP test shows which features have the biggest impact on the model's choices. 'IMU_ankle_17' and 'IMU_hand_1' are very important. They show that putting sensors on the ankle and hand gives important information for sorting activities. This knowledge can help plan future data-gathering tactics, so sensors are placed well and data quality is focused on for these readings. Also, it would be helpful to look at how these important features relate to each other. This could give us more information that might improve how well the model works, especially when it's not very good at predicting certain classes.

Model Interpretability and Trust: Analyzed SHAP values, which explain the impact of each data point on the model’s predictions. This helps understanding what influences the model most when it's identifying activities. This makes it easier to understand, and it also helps people involved in the project feel more confident. This openness is important when using the model in real-life situations. That's because people and workers depend on what the system suggests to make decisions.

10. Recommendations

Refined Data Collection for Underperforming Classes:

Look into how data is collected for activities related to Classes 10, 11, and 12. We might need to use smaller sensors or add extra types of sensors to get all the movements involved in these activities.

Advanced Feature Engineering:

Create new parts that might show the hardness of real actions better. This could include deeper connections between existing features, checking the data over time or patterns in sounds that may show patterns not found in the normal data.

Model Ensembles and Advanced Algorithms:

Try using group models that mix the guesses from many machine learning systems. Think about fancy methods like deep learning, which might be better at finding complicated patterns in the data.

User-Centric Model Tuning:

Make the model suit each user, maybe making special models for one person that considers their special body parts or what they usually do.

Stakeholder Feedback Integration:

Talk to the people who will use the model and experts in the field to get their thoughts on how well it works and how easy it is to use. Their knowledge could show useful thoughts that might not be seen from the information alone.

Deployment and Real-time Learning:

Put the model in a real-life setting where it can learn from the constant flow of data. This will make the model better over time and make sure it's still important when new types of activities come up.

Robustness and Edge Case Analysis:

Check the model's strength by doing a stress test. Look at how well the model works on unusual cases and rare actions, and create plans to deal with these situations. This could be done by creating special models or by using rule-based systems.