Received: 10 June 2024 | Revised: 18 December 2024 | Accepted: 19 February 2025 | Published online: 11 March 2025

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2025, Vol. 00(00) 1–12
DOI: 10.47852/bonviewJDSIS52023588

BON VIEW PUBLISHING

Comprehensive Health Tracking Through Machine Learning and Wearable Technology

Abusufyan Yusuf[1], Tareq Al Jaber[2] and Neil Gordon[2,*]

[1]*Faculty of Science and Engineering, University of Hull, UK*

[2]*School of Computer Science, University of Hull, UK*

Abstract: The accurate interpretation of data from wearable devices is paramount in advancing personalized healthcare and disease prevention. This study explores the application of machine learning techniques to improve the interpretation of health metrics from wearable technology, focusing on heart rate and activity prediction. The study conducts a device-wise comparison of data from popular devices, namely the Apple Watch and Fitbit, using both tree-based and boosting algorithms. The outcome of the experiment shows that the Random Forest model is a better predictor for heart rate, with the lowest error rate across devices and a prediction accuracy of 98% on the combined dataset. Conversely, the classification result for activity prediction showed that all models used have better accuracy with Fitbit data, and accuracy drops with Apple Watch data. The Random Forest achieves a consistent performance of 87% for accuracy and $F1$ score on the combined data. However, after cross-validated hyperparameter tuning, this result on the combined dataset is superseded by the boosted models, with both Gradient Boosting and XGBoost achieving the same level of performance (90%) across metrics.

Keywords: wearable device technology, health tracking, machine learning, activity prediction, data science

1. Introduction

Health tracking has evolved into an indispensable tool for enhancing overall wellness and preventing diseases, as evidenced by the studies conducted by Prieto-Avalos et al. [1], Stonjancic et al. [2], and Sethi et al. [3]. The emergence of wearable devices equipped with embedded sensors has revolutionized health monitoring, allowing for real-time tracking of a diverse range of physiological and physical parameters, as highlighted by Tang et al. [4]. This technological advancement empowers individuals to gain a comprehensive understanding of their health status. The sensors integrated into wearable devices capture a broad spectrum of data, ranging from heart rate to calorie consumption, providing users with instant feedback crucial for proactive health management, as indicated by the research of Hussain et al. [5] and Nnaji et al. [6].

Despite the exponential growth in the use of these health-tracking devices, a significant challenge arises in effectively harnessing the vast amount of data they generate. Recognizing this challenge, this project aims to go beyond the mere utilization of health-tracking data. Instead, it focuses on extracting valuable insights through the examination of advanced activity categorization and heart rate prediction, utilizing the latest machine learning approaches [7] to offer insights into the data, in contrast to existing approaches that typically focus on singular aspects of prediction. Multiple algorithms will be employed across various device profiles to enhance the accuracy and reliability of the predictions by identifying the most effective algorithms for analyzing this data.

In contrast to previous works that typically concentrate on a singular aspect of prediction, often centered around activity categorization, this project takes a more comprehensive approach. By addressing multiple dimensions of health tracking, it seeks to contribute to the existing body of knowledge and advance our understanding of how different algorithms perform across various health parameters. This research endeavor aligns with the increasing importance of utilizing technology to not only monitor health but also to derive actionable insight for personalized and effective health management.

The significance of this project is underscored by the growing ubiquity of wearable devices, as reported by Bloomberg [8], making it imperative to explore innovative ways to leverage the data they generate. The outcomes of this study have the potential to inform the development of more sophisticated and accurate health-tracking algorithms, thereby enhancing the utility of wearable devices in promoting individual well-being

1.1. Aims and objectives

The primary aim of this project is to derive advanced insights into the reliability and functionality of wearable devices based on the data they produce. The objective is dissected into two core tasks, each complemented by specific sub-tasks, as detailed below:

*Corresponding author: Neil Gordon, School of Computer Science, University of Hull, UK. Email: N.A.Gordon@hull.ac.uk

1.1.1. Data collection and preliminary analysis

1) Import Data: Acquire and consolidate data from two wearable devices to establish a comprehensive dataset for analysis.
2) Check Statistical Distribution: Ensure that the data is representative and without significant biases by analyzing its statistical properties.
3) Exploratory Data Analysis (EDA): Conduct an initial exploration of the data to identify patterns and potential areas of interest.
4) Identify Gender from Attributes: Analyze the given attributes to deduce the gender of the wearable device user.
5) Device Profiling: Categorize the readings from the devices to understand their variability and potential reliability.

1.1.2. Advanced analysis and model implementation

1) Multiple Classifiers for Activities: Use machine learning classifiers to predict and categorize the types of activities based on data attributes.
2) Multiple Regressors for Heart Rate: Utilize regression models to predict the heart rate of participants based on the data from wearable devices.

1.2. Research questions

1) This contribution in this work is shown through the following research questions that are answered later:
2) Can a unified, multi-output model effectively leverage shared patterns in the data to achieve comparable or superior predictive performance for both heart rate and activity, as compared to models trained individually for each target?
3) To what extent does the predictive accuracy of machine learning models for heart rate and activity classification depend on the type of wearable device used?
4) Is there any evidence of anomalies in the dataset?
5) Can probabilistic hyperparameter tuning improve model performance on the combined data?

2. Literature Review

Historically, research in the realm of wearable devices has concentrated on the reliability and validity of these devices, focusing mainly on step counts, heart rate, and energy expenditure. While valuable, these efforts often do not tap into the potential of advanced data analytics or deep learning techniques. Several studies have evaluated the accuracy of wearable devices in monitoring basic physical activities. In 2017, Sztyler et al. [9] developed a system using a Random Forest Classifier to enhance wearable-based activity recognition, aiming to accurately detect both the on-body position of wearable devices and the type of physical activity being performed, even when device placement varies across users. Their research underscores the significance of device localization in improving recognition rates, especially when dealing with activities like standing, lying, and sitting, which might be otherwise confused due to their similar low acceleration profiles. The results revealed that they achieved an 81% success rate in identifying device locations during different activities, and the shin proved to be a more reliable position than the forearm.

In the study by Oyeleye et al. [10], the research focused on predicting heart rate using data from wearable accelerometers, with the goal of early detection and management of heart disease. The study tested various models including the ARIMA model, Linear Regression, Support Vector Regression, K-Nearest Neighbor Regressor, Decision Tree Regressor, Random Forest Regressor, and Long Short-Term Memory Recurrent Neural Network, on a newly created dataset. The results showed that the ARIMA model, particularly when combined with walk-forward validation and linear regression, was effective in predicting heart rate for all durations of activity better.

Manjarres et al. [11] explored the integration of human activity recognition with heart rate measurements to calculate workload, targeting applications in workplace health and fitness. The system utilized a traditional machine learning algorithm, specifically a Random Forest classifier, to discern user activities, achieving an accuracy of up to 92%. A focused case study further illustrated the systems' capability to monitor physical workload adaptation over a span of twenty days. This framework aimed to streamline workload monitoring for ergonomics and health professionals.

Similarly, Choksatchawathi et al. [12] conducted a study addressing the accuracy of heart rate measurements from wearable devices in varying states of daily activity, recognizing the challenges posed by motion sensors in capturing accurate measurement during activities. They tested four popular wearable devices in a study with 29 participants, covering different activities such as resting, laying down, and intense treadmill exercise. The study revealed high error rates in heart rate measurements, especially from the Fitbit Charge HR. By developing an improved heart rate estimation model using rolling windows as a feature, they significantly reduced the mean absolute error in measurements, demonstrating the feasibility of enhancing heart rate monitoring accuracy in daily use.

Utilizing the same dataset as our study, Fuller et al. [13] conducted a study to evaluate the accuracy of wearable devices in calibrating various activities. They employed a range of classification models, including Decision Trees, Support Vector Machines, and Random Forest algorithms, with the aim of ascertaining the overall percentage accuracy of these wearables. The outcomes demonstrated variability across different devices. However, the Random Forest algorithm exhibited 82% accuracy for the Apple Watch and an even higher 90% for the Fitbit Watch.

More recent work [14, 15] has applied machine learning using more advanced wearable technology, whereas this study considers readying available domestic devices.

This project seeks to build on the existing foundation by incorporating activity recognition with heart rate detection with an experimental effort in exploring how distinct and combined wearable device data can influence the accuracy and robustness of health monitoring systems.

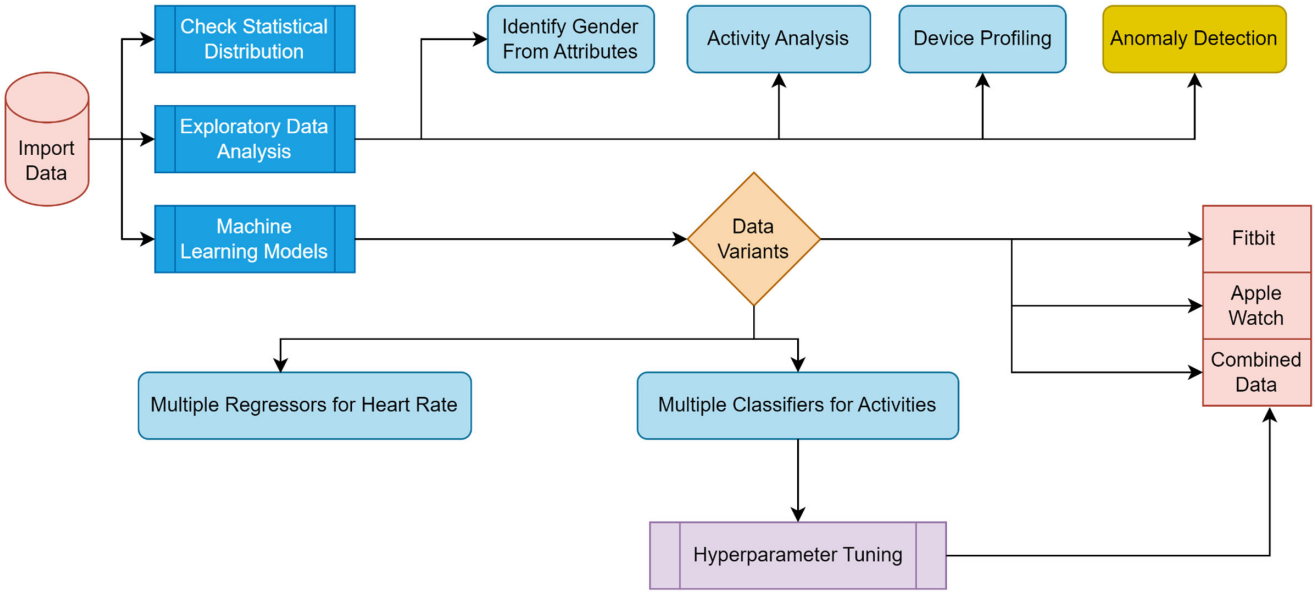
3. Research Methodology

Figure 1 shows the intended approach of the proposed implementation. This starts with the import of data, then the EDA, with gender, activity, and device profiling, and anomaly detection. The machine learning goes through the data variants, where it branches based on the devices and with classifiers for activities, as well as the heart rate analysis.

3.1. Data overview

The dataset selected for this study was retrieved from a public repository and originated from a study conducted to investigate the accuracy and efficacy of activity tracking in varied real-world scenarios [12]. The data collection experiment involved participants wearing multiple wearable devices, using Apple Watch Series 2 and Fitbit Charge HR2, where they were subjected to a structured 65-minute protocol encompassing various activities

Figure 1
Conceptual framework of the proposed implementation



ranging from sedentary behaviors like lying and sitting to more dynamic activities like walking and running. The collected data spans eighteen variables, including heart rate, steps taken, calories burned, and distance traveled, among others (Table 1). The combined dataset is a total of six thousand two hundred sixty-four data points (3656 from Apple Watch and 2608 from Fitbit Watch), with 52% representing female observations and the remainder for male counterparts.

Table 1
Dataset variable names and description

S/N	Variables	Description
1	Age	Age of participants
2	Gender	Gender of participants
3	Height	Height of participants
4	Weight	Weight of participants
5	Steps	Number of steps/minute
6	Heart Rate	Average heart rate/minute
7	Calories	Amount of calories expended
8	Distance	Distance ran in meters
9	Entropy Heart	Measure of heart rate variability
10	Entropy Steps	Measure of steps variability
11	Resting Heart	10th percentile of HR data
12	Corr heart steps	Correlation coefficient between heart rate and steps
13	Norm Heart	Normalized heart rate
14	Intensity Karvonen	Intensity zone during activity
15	Sd Norm Heart	Standard deviation of normalized heart rate
16	Steps times Distance	Total amount of steps and distance covered in meters
17	Device	Device type
18	Activity	Activity engaged in

3.2. Exploratory Data Analysis (EDA)

3.2.1. Gender identification

One of the unique challenges faced during the preliminary exploration of the dataset was the pre-processed nature of the gender attribute, which made it difficult to ascertain the gender classification. To resolve this, we leveraged specific physiological parameters such as height, weight, and resting heart rate to analyze and decode the gender classes.

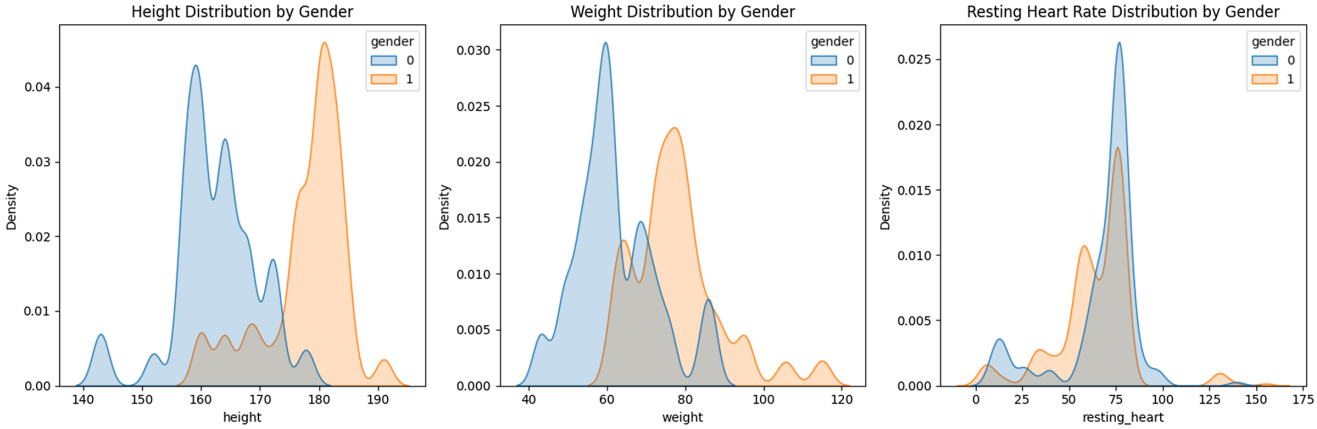
Height: Group labeled 0 had an average height of approximately 162.46 cm, ranging from 143.0 cm to a maximum of 177.8 cm. Group 1, on the other hand, displayed an average height of around 177.67 cm, ranging between 160.0 cm and 191.0 cm.

Weight: The average weight for the group 0 was observed to be around 62.17 kg, with individuals weighing as little as 43.0 kg and as much as 86.4 kg. For group 1, the average weight stood at approximately 77.8 kg, with a spectrum spanning from 62.0 kg to 115.0 kg.

Resting Heart Rate: Participants in group 0 exhibited an average resting heart rate of approximately 67.53 bpm, with the rate fluctuating between 10.0 bpm and 140.0 bpm. Conversely, group 1 had an average resting heart rate of around 64.04 bpm, with individual rates ranging from a mere 3.0 bpm to 155.0 bpm. These ranges exhibit abnormal values, indicative of issues with the values reported by the devices. Given the focus of this work on the mobile devices, these have been left in. For diagnostic and treatment purposes such values would need to be removed through anomaly data filtering as part of the data-preprocessing.

From the aforementioned parameters and considering established physiological differences between typical male and female attributes [16, 17], it became evident that group 0 represented females while group 1 denoted males. This is

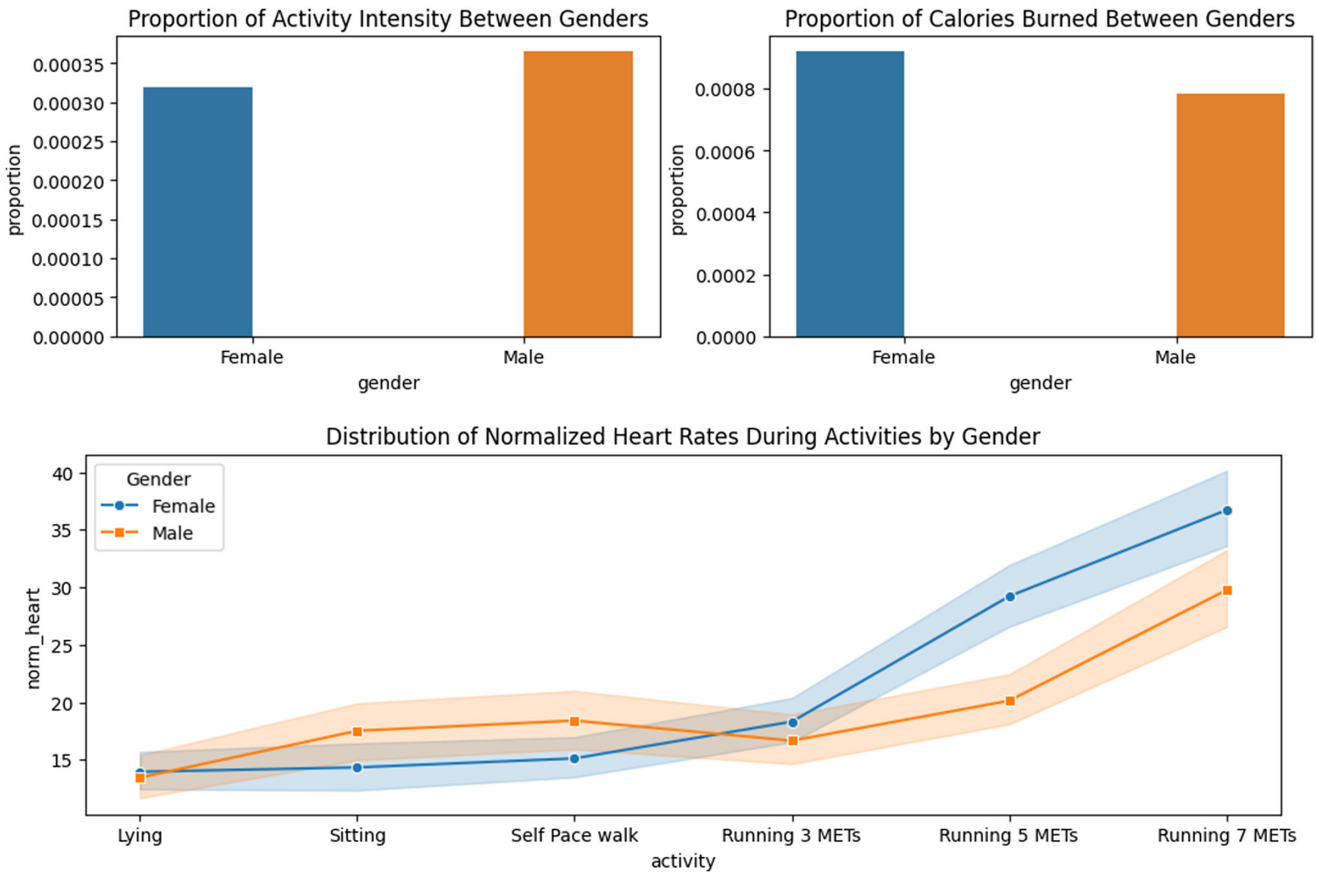
Figure 2
Physiological parameters in identifying gender



illustrated in Figure 2, where the graphs show how different genders have different profiles.

Figure 3 shows the physiological differences between male and female participants during various physical activities. The top two bar charts show the proportional activity intensity and calories burned by each gender. The bottom-line plot shows the distribution of normalized heart rates for both genders across the activities recorded, from passive states like lying and sitting to more intense exertions such as self-paced walking and running at multiple MET levels. The shaded areas around the lines indicate the variability in heart rates. A key observation is that women generally have higher normalized heart rates across these activities

Figure 3
Physiological differences between male and female engaged in the same activities



compared to men, which explains why women burned more calories in the observed activities compared to men who engaged in more intense activities.

3.2.2. Device profiling

Device profiling was carried out as a technique to measure the reliability of the devices used to capture the data. The hypothesis behind this exercise is that both devices—Apple Watch and Fitbit—would exhibit similar performance and accuracy in tracking metrics across various activities.

The outcome from Table 2 shows that Apple Watch has greater variability in calorie tracking, with a higher standard deviation than the Fitbit. However, its consistency was more reliable in step counting, as indicated by a lower standard deviation. Both devices showed similar variability in heart rate monitoring, with the Apple Watch being slightly more consistent. In contrast, the Fitbit exhibited substantial variability in distance tracking, as reflected by a much higher standard deviation compared to the Apple Watch. These findings suggest that the Apple Watch may provide more consistent step counts and slightly more stable heart rate measurements, while the Fitbit may offer more consistent calorie tracking but less reliable distance measurements. The implications of these variations for users who rely on accurate health metrics are multifaceted, depending on the users' specific health goals and the activities they engage in. For instance, individuals who prioritize precise calorie tracking for diet and weight management might lean towards the Fitbit, given its lower standard deviation in calorie measurements. Conversely, for those who focus on step counting, such as walking or running enthusiasts, the Apple Watch may be more beneficial due to its lower variability in steps tracking.

Table 2
Intra-device consistency (Standard deviation)

Device	Calories	Heart rate	Steps	Distance
Apple Watch	269.438	26.753	7.279	0.137
Fitbit	25.344	26.671	32.961	66.587

3.3. Model development

The model development involves a comprehensive exploration of various predictive models to analyze the wearable device data. The process begins with establishing a base using a Multi Output Regressor, which extends the capabilities of a Random Forest Regressor to predict multiple target (heart rate and activity) variables simultaneously. The model achieves its multi-target predictive power by constructing separate instances for each response variable.

The Multi Output Regressor works as follows:

For a training set with n samples, $X = \{x1, x2 \ldots, xn\}$ is the matrix of input features and $Y = \{y1, y2, \ldots, yn\}$ is the matrix of the target variables, where each $y1$ is a vector containing the values of the heart rate and activity for the $i - th$ sample. The Multi Output Regressor fits one Random Forest Regressor for each target variable using the input matrix corresponding to the target vector.

To address the central hypothesis, the model development was advanced with the implementation of multiple machine learning algorithms, each selected for their proven abilities in pattern detection and predictive accuracy. These algorithms were leveraged in both their classifier and regressor forms to maintain uniformity and facilitate direct comparison between categorical and continuous predictions.

Tree-Based Models

The Random Forest and Decision Tree models used fall within this category and make predictions based on a hierarchy of binary splits (called trees) in the features [18].

Decision Tree: Creates a flowchart-like structure, where each internal node represents a decision on a feature, each branch represents an outcome of the decision, and each leaf node represents a final classification or regression value. The classifier version assigns a class label to each leaf, leading to categorical predictions. The regressor version predicts continuous values, using the mean or median of the target variable in each leaf node.

$$\hat{Y} = \begin{cases} mode(Y_{leaf\ mode}) \ for\ classifier \\ mean(Y_{leaf\ mode}) \ for\ regressor \end{cases}.$$

where:
$\hat{Y}$ = the predicted outcome
$Y_{leaf\ node}$ = the set of data points in a leaf node of the tree.
mode = the most frequent label in the case of classification.
mean = the average of values for regression.

Random Forest: Builds upon the concept of a Decision Tree but incorporates multiple trees to improve prediction accuracy and control overfitting [19]. In its classifier form, it aggregates the predictions from numerous decision trees to decide on the final class, effectively reducing the risk of errors from any single tree [20]. As a regressor, the Random Forest takes the average of all the decision trees' predictions, providing a more reliable and robust estimate for continuous variables [21]. This ensemble approach, where multiple models contribute to the final decision, enhances the model's performance.

$$\hat{Y} = \frac{1}{K}\sum_{k=1}^{K} \begin{cases} T_k^{\text{mode}}(X) \ for\ classifier \\ T_k^{\text{mean}}(X) \ \ for\ regressor \end{cases}$$

where:
$\hat{Y}$ = the predicted outcome
K = the number of trees in the forest.
T_k = is the prediction of the k^{th} tree.
X = the input feature set.

Boosting Models

Boosting models are characterized by their ability to build strong predictive models by sequentially combining weaker ones, often decision trees [22].

Gradient Boosting:

The classifier variant focuses on minimizing misclassification errors by building sequential trees, where each tree corrects the mistakes made by the previous one [23]. In the regression context, it works on the same principle but targets continuous data, minimizing the Mean Squared Error.

$$F_M(\mathrm{X}) = \sum_{m=1}^{M} \alpha_m f_m(X)$$

where:
F_M (X) = the final model after M boosting rounds.

α_m = the learning rate for the mth model
f_m = is the output of the m^{th} weak learner.

XGBoost: Extreme Gradient Boosting (XGBoost) further refines the concept of gradient boosting by incorporating regularization techniques, which prevents overfitting. In both instances, it emphasizes on optimizing computational efficiency and model performance by maximizing predictive accuracy while managing computational resources effectively [24].

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where:
Obj = the objective function to be minimized.
$L(y_i\text{-}\hat{y}_i)$ = the loss function, comparing the actual value and the predicted value.
$\Omega(f_k)$ = the regularization term for the k^{th} tree, controlling model complexity.

In the next phase of model development, all classifier variants were subjected to a hyperparameter tuning process using a probabilistic approach. This method explored RandomizedSearchCV, which integrates cross-validation and random selection of combinations of hyperparameters to identify optimal settings [25, 26]. This probability ensures an unbiased tuning process, for fine-tuning the classifiers to the specifics of our dataset while maintaining their ability to generalize to new data.

4. Results

Tables 3 and 4 show the results of the device-wise exploration for the four models used in the regression and classification analyses. The regression models were evaluated using the mean absolute error (MAE) and the coefficient of determination (R^2), which provide a clear picture of the models' prediction accuracy and the variance they captured from the dataset. This is illustrated in Figure 4. The fit of the models to the data was also visualized through lines of best fit plotted against the observed values (see Figure 5).

The Gradient Boosting, XGBoost, and Random Forest models achieve high R^2 scores close to 100% across the devices (Table 3), suggesting an excellent fit to the data. However, the Random Forest model resulted in the lowest MAE in all instances, with high R^2 scores, indicating precise predictions with minimal average error.

The base model achieved the highest R^2 score and lowest MAE, making it a better model for predicting heart rate; however, its performance on the classification metrics (Table 4) was suboptimal, suggesting that while it can predict numerical values effectively, it struggles with categorizing data, which is essential for a multi-output model's overall accuracy and utility.

The classification models were evaluated using accuracy and *F*1 score to ascertain their effectiveness in correctly identifying categories. To complement these metrics and offer a deeper insight into the classification performance, confusion matrices were provided for each model, illustrating the distribution of true positives, false negatives, true negatives, and false positives. This is illustrated in Figure 6, with the classification analysis by device, and Figure 7 summarizes the confusion matrices.

XGBoost and Random Forest exhibit superior performance across both individual devices and the combined dataset, achieving the highest accuracy and *F*1 scores in activity classification. The Gradient Boosting and Decision Tree models demonstrate moderate effectiveness. These results lead to the implementation of a hyperparameter tuning technique to further enhance the classification results.

On average, without tuned hyperparameters, the comparison of correctly classified activities across the models indicates that the

Table 3
Outcome of device-wise regression analysis to predict heart rate

Device	Fitbit		Apple Watch		Combined	
Models	R2 score	MAE	R2 score	MAE	R2 score	MAE
Base Model	1.00	0.809	1.00	0.59	1.00	0.54
Gradient Boosting	0.997	0.809	0.998	0.815	0.996	1.077
XGBoost	0.997	0.508	0.998	0.659	0.997	0.613
Random Forest	0.996	0.457	0.998	0.595	0.998	0.538
Decision Tree	0.995	0.558	0.994	1.103	0.996	0.798

Table 4
Outcome of a device-wise experiment to predict activities

Device	Fitbit		Apple Watch		Combined	
Models	Accuracy	*F*1 score	Accuracy	*F*1 score	Accuracy	*F*1 score
Base Model	0.80	0.81	0.54	0.51	0.62	0.61
Gradient Boosting	0.883	0.883	0.724	0.723	0.740	0.740
XGBoost	0.902	0.902	0.855	0.855	0.865	0.866
Random Forest	0.900	0.900	0.844	0.843	0.878	0.878
Decision Tree	0.871	0.872	0.745	0.746	0.776	0.776
Benchmark on the same dataset [12]						
Random Forest	0.9080	–	0.8195	–	–	–
Decision Tree	0.6234	–	0.4139	–	–	–

Figure 4
Visual comparison of device-wise result for regression analysis

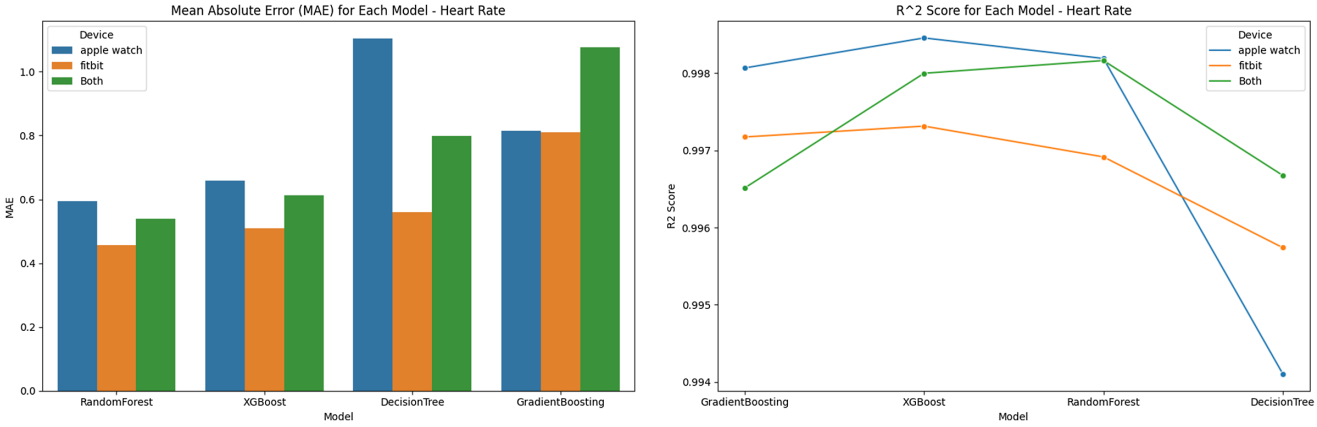


Figure 5
Comparison of line of best fit across regression models

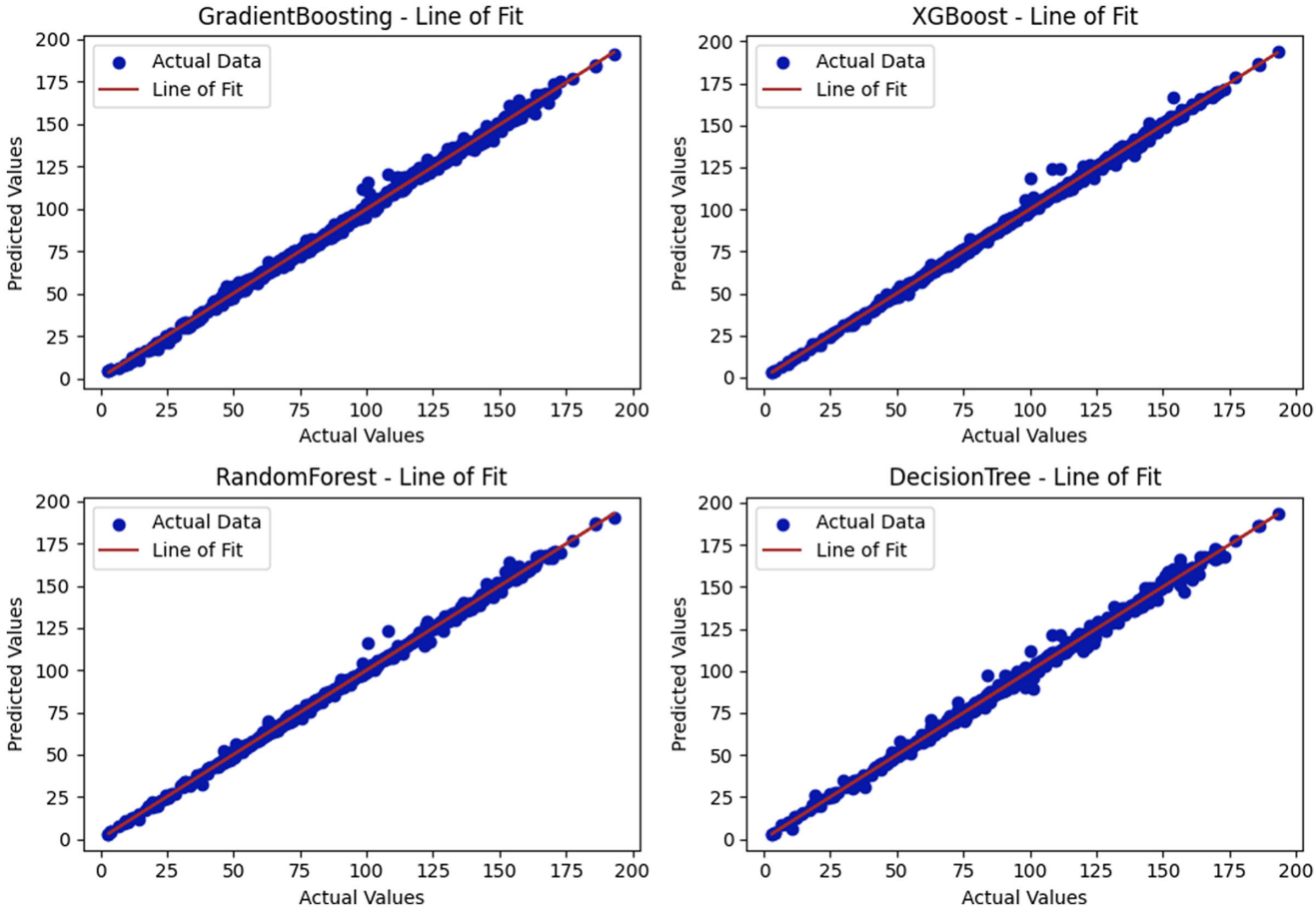
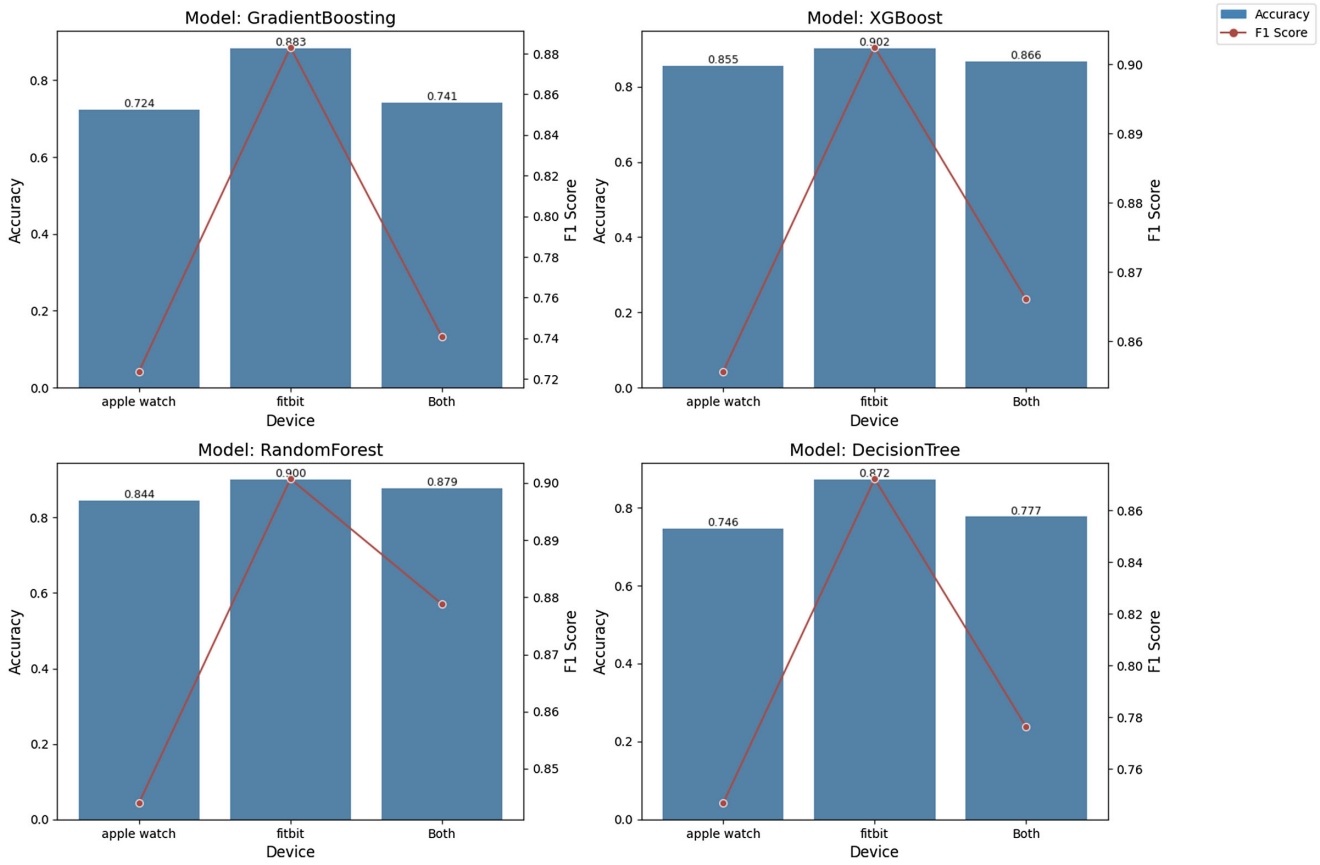


Figure 6
Visual comparison of device-wise result for classification analysis



random forest model makes a minimal error in misclassifying activities in the dataset.

Figure 8 compares the probabilistic optimization result with the models before tuning. The Random Forest classifier was the top-performing model before tuning. However, Gradient Boosting has taken the lead after tuning, with improvements in both metrics. This indicates a more balanced performance in terms of precision and recall. The Tree-based classifiers, despite improvements after tuning, do not outperform the boosted algorithms.

We analyzed heart rate data to assess the reliability of wearable device data in monitoring user health. We hypothesized that anomalous readings could indicate unusual physical responses in data capture, which might be important for user health monitoring and intervention. We calculated the Z-score for each heart rate reading against the number of steps taken during various activities to determine the presence of anomalies. The mathematical foundation of the Z-score test is based on the standardization of data points:

$$Z = \frac{(x - \mu)}{\sigma}$$

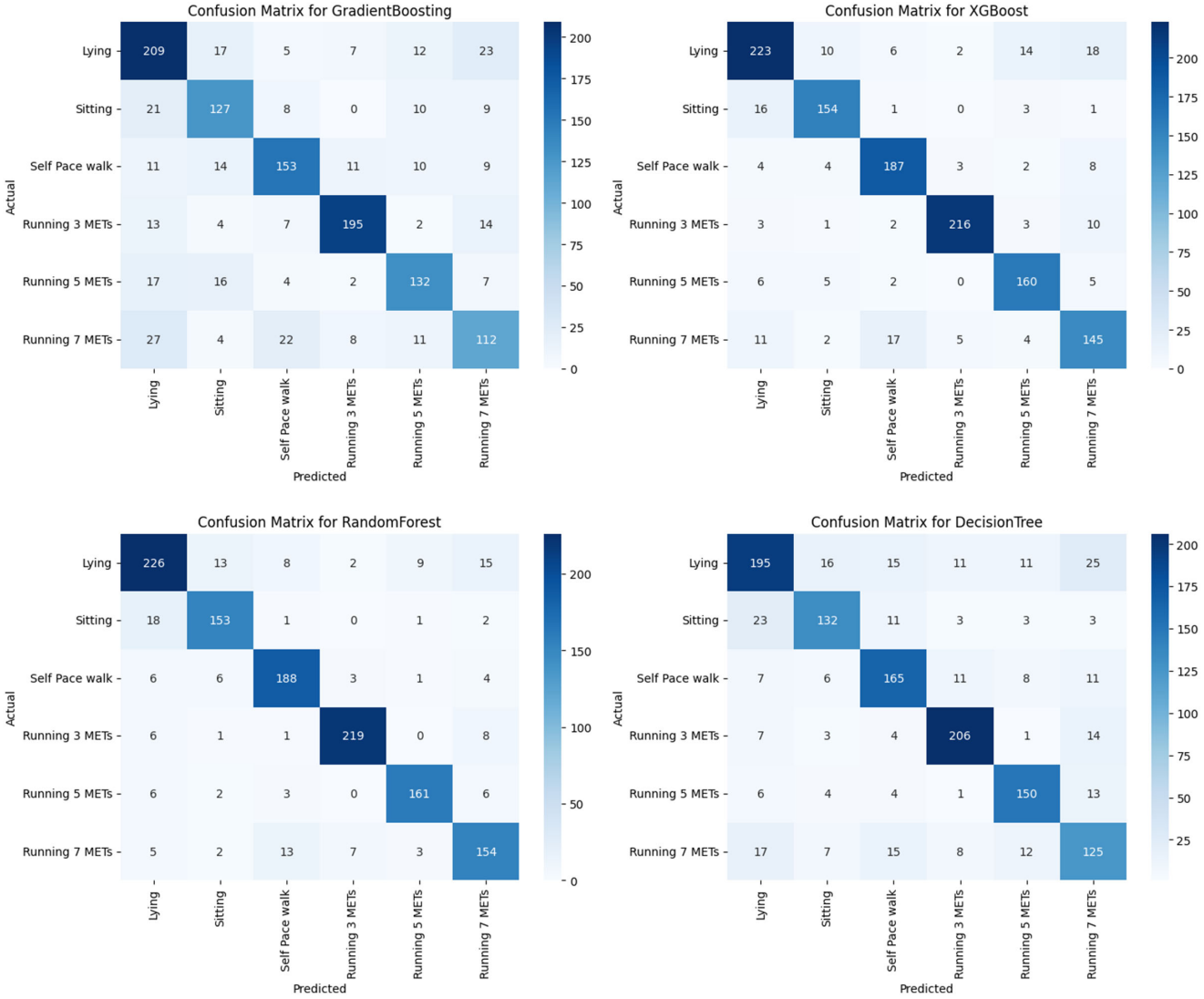
where:
Z = the z-score of the data point.
x = the value of the data point.
μ = the mean of the dataset.
σ = the standard deviation of the dataset.

Using a threshold of ±2 for the z-score, Figure 9 indicates that most heart rate values fall within expected ranges during active and inactive states, as shown by the blue points. However, there are minor red points that stand out during low-intensity activities such as lying and sitting, which are not typically associated with elevated heart rates. This could signal discrepancies that may be due to physiological variations among individuals, anomalies in measurement, or misclassification of activity types. During higher-intensity activities, such as running, the presence of red points suggests sporadic deviations from expected heart rate values, which also warrants further examination to discern between true physiological stress and potential health issues for the users.

5. Discussion

The base model's performance in predicting heart rate with an R^2 score of 1.00 and the lowest MAE when trained on the combined dataset, paradoxically fell short in the classification analysis. This highlights that there's an inherent complexity in crafting multi-output models that are proficient at both regression and classification tasks. From the device-wise analysis, the classifiers exhibited a poorer performance for the Apple Watch data, despite it constituting a larger proportion of the dataset with 3656 instances compared to the Fitbit's 2608. This warrants a closer examination of the models' ability to generalize across varying data distributions and device measurement calibrations. The Apple Watch's poorer performance signals that there are patterns within its data, that the models struggle to accurately interpret, and this is

Figure 7
Comparison of confusion matrices across classification models



why the accuracy of the classification models before tuning on the combined dataset drops and the error rate increases.

The benchmarked study by Fuller et al. [12] focused on device-wise analysis and achieved 81% accuracy for the Apple Watch and 90% for the Fitbit using RandomForest models. Our analysis, in contrast, revealed that our RandomForest model matched their 90% accuracy for Fitbit and exceeded it for Apple Watch with 3% increase. Similarly, our Decision Tree models outperformed the benchmarks, registering a significant 25% improvement for Fitbit and an even more impressive 33% enhancement for Apple Watch. Post-hyperparameter tuning, our boosting models demonstrated significant improvements, achieving a balanced 90% accuracy and F1 score on the combined dataset.

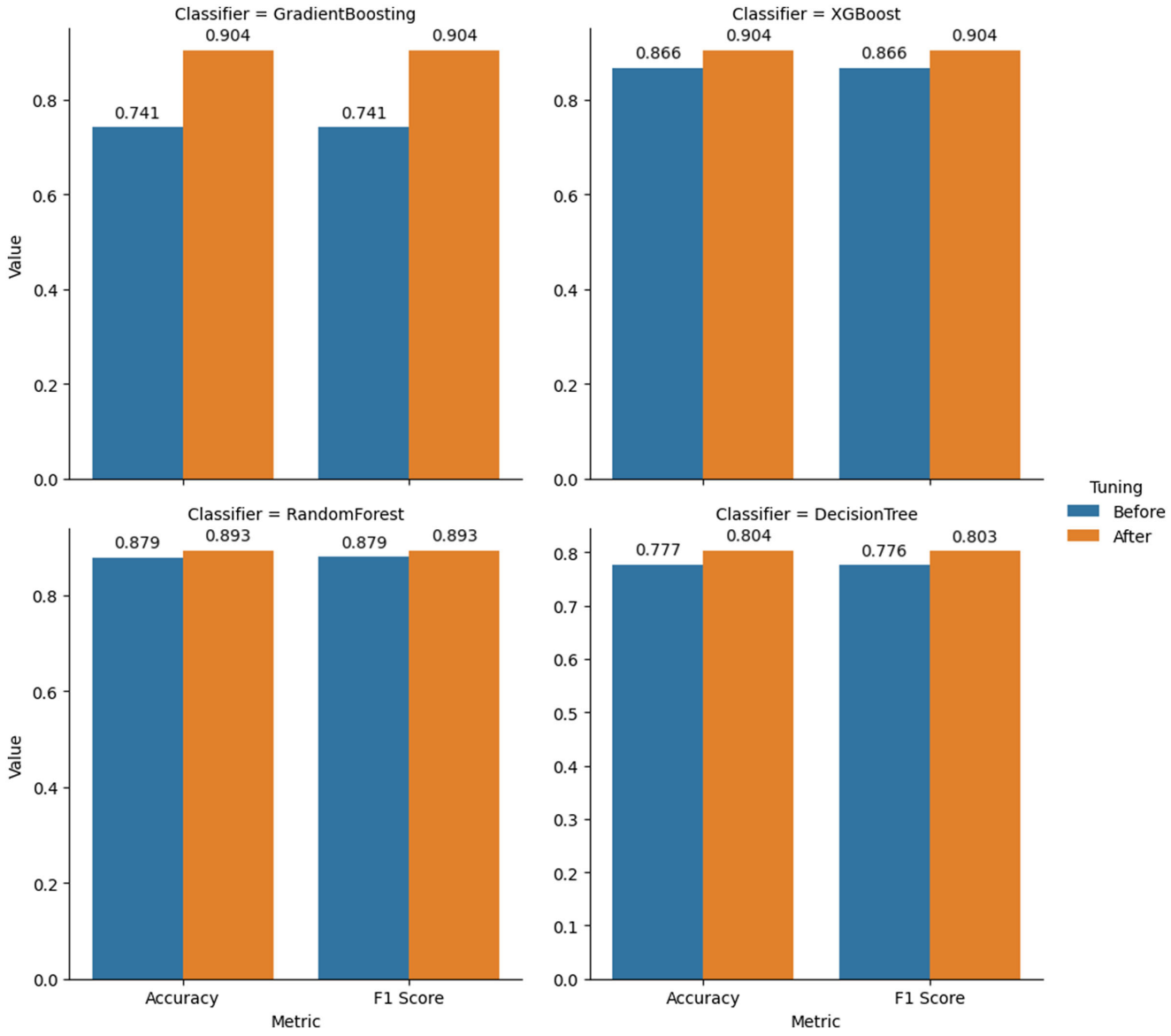
This analysis shows that, in terms of the original research questions

1) A unified, multi-output model can effectively demonstrate patterns in the data to improve predictive performance
2) the predictive accuracy of machine learning models for heart rate and activity classification depends on the type of wearable device used
3) As identified in Section 4, there is some evidence of anomalies in the dataset which would require addressing for in actual health applications
4) Probabilistic hyperparameter tuning can improve model performance.

5.1. Limitations

Due to data availability, our study was limited to two wearable devices. The models may not perform as well with data from other devices. Different devices use different sensors and algorithms, which can lead to discrepancies in data quality and format. Our findings may not be universally applicable to all wearable health technologies. The reported work is comprehensive in considering the readily available data and platforms, but these are somewhat

Figure 8
Comparison of classification models before and after probabilistic optimization



limited and a broader set of measurements, such as blood pressure would have been helpful.

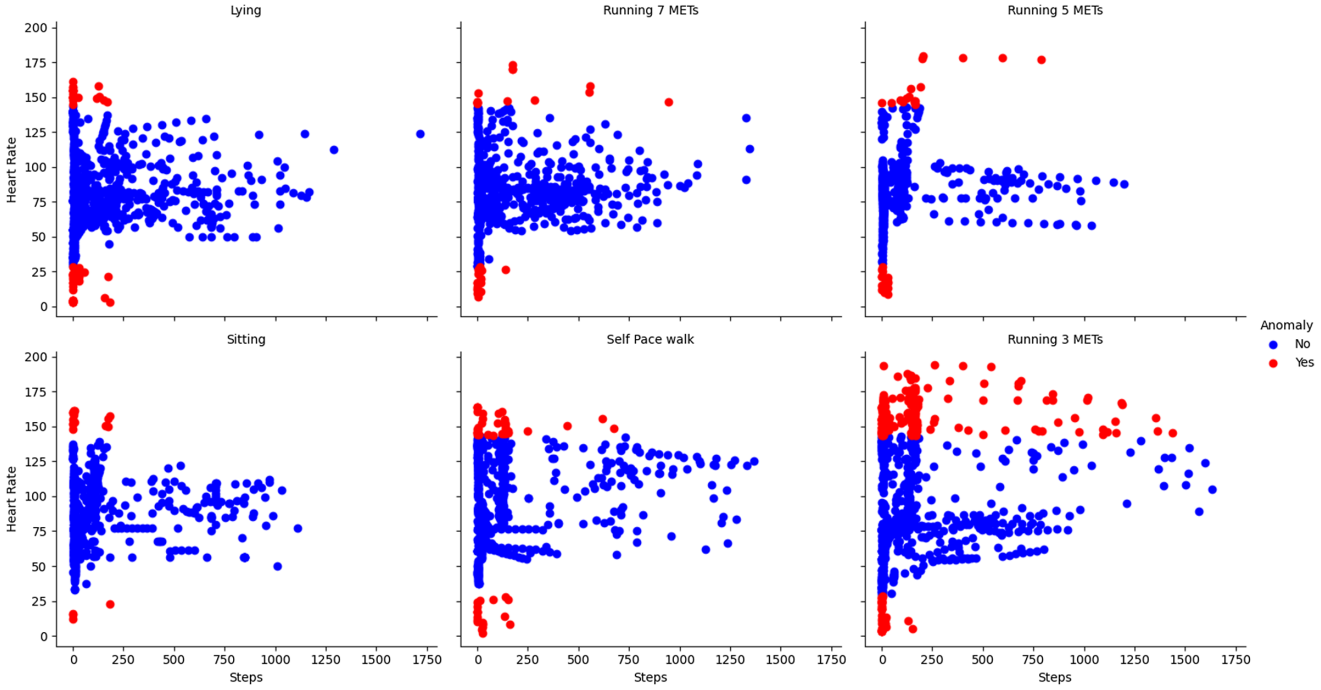
Another limitation is the potential exclusion of external factors that could impact heart rate readings and activity classification. Our models were trained and validated on datasets that may not fully encompass the effects of individual health conditions, medication intake, stress levels, and other environmental variables. A comprehensive model would need to account for these factors.

5.2. Future works

To build upon the current study, future research initiatives should prioritize the expansion of the dataset to include a broader spectrum of devices to capture a more heterogeneous sample population. Collecting data across diverse demographics would bolster the robustness of the models. Additionally, gathering data under varied conditions such as in fluctuating environmental settings, and across multiple geographic locations, would significantly enrich the dataset.

To truly assess the practicality and reliability of these models, it is beneficial to test them in real-world scenarios. Future studies should focus on integrating these models into everyday health monitoring systems to their performance in dynamic, uncontrolled settings. This would not only validate the utility of the models in practical applications but also highlight areas that require further refinement.

Figure 9
Anomaly detection using Z-Score on activity levels



6. Conclusions

This comprehensive study on the application of machine learning techniques and wearable technology for health tracking has yielded remarkable insights. The extensive analysis of wearable device data for heart rate prediction and activity categorization reemphasizes the potential of these technologies in personal health management. The findings from device profiling reveal that while both the Apple Watch and Fitbit have their strengths, each device's data accuracy varies depending on the health metric in question. The high predictive accuracy of the machine learning models, particularly after probabilistic hyperparameter tuning for activity classification, suggests that wearable technology can be a reliable tool for health monitoring and potentially for disease prevention. However, the variability in model performance across different devices highlights the need for further research to optimize these technologies for more accurate and personalized health tracking.

Acknowledgements

The authors acknowledge Dr Habeeb Balogun of Big Data Technologies and Innovation Lab, University of Hertfordshire, UK, who provides constructive criticism for this work. Additionally, the authors also acknowledge the Harvard Dataverse publisher for providing data used for this research.

Finally, we wish to thank the reviewers for constructive comments and feedback that led to further revisions and enhancements of this paper.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors. The work was carried out according to the ethical requirements of the University of Hull.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study. The code—and parameters for the different algorithms—can be made available by contacting the authors.

Author Contribution Statement

Abusufyan Yusuf: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Tareq Al Jaber:** Conceptualization, Data curation, Writing – review & editing, Supervision. **Neil Gordon:** Conceptualization, Data curation, Writing – review & editing.

References

[1] Prieto-Avalos, G., Cruz-Ramos, N. A., Alor-Hernández, G., Sánchez-Cervantes, J. L., Rodríguez-Mazahua, L., & Guarneros-Nolasco, L. R. (2022). Wearable devices for physical monitoring of heart: A review. *Biosensors*, *12*(5), 292.

[2] Stojancic, R. S., Subramaniam, A., Vuong, C., Utkarsh, K., Golbasi, N., Fernandez, O., & Shah, N. (2023). Predicting pain in people with sickle cell disease in the day hospital using the commercial wearable apple watch: Feasibility study. *JMIR Formative Research*, *7*, e45355.

[3] Sethi, V., Katal, A., Dabas, S., & Kumar, S. (2022). Heart attack detector: An IoT based solution integrated with cloud. In *2022 13th International Conference on Computing Communication and Networking Technologies*, 1–5.

[4] Tang, J., El Atrache, R., Yu, S., Asif, U., Jackson, M., Roy, S., & Loddenkemper, T. (2021). Seizure detection using wearable sensors and machine learning: Setting a benchmark. *Epilepsia*, *62*(8), 1807–1819.

[5] Hussain, G., Ali Saleh Al-rimy, B., Hussain, S., Albarrak, A. M., Qasem, S. N., & Ali, Z. (2022). Smart piezoelectric-based wearable system for calorie intake estimation using machine learning. *Applied Sciences*, *12*(12), 6135.

[6] Nnaji, C., Awolusi, I., Park, J., & Albert, A. (2021). Wearable sensing devices: Towards the development of a personalized system for construction safety and health risk mitigation. *Sensors*, *21*(3), 682.

[7] Saleem, I., & Irfan, A. (2024). Machine learning made easy: A beginner's guide for causal inference and discovery methods using python. *International Journal of Data Analysis Techniques and Strategies*, *17*. https://doi.org/10.1504/IJDATS.2025.10064732

[8] PR Newswire. (2022). *Wearable technology market to hit $186.14 billion by 2030: Grand View Research, Inc.* Retrieved from: https://www.prnewswire.com/news-releases/wearable-technology-market-to-hit-186-14-billion-by-2030-grand-view-research-inc-301687628.html

[9] Sztyler, T., Stuckenschmidt, H., & Petrich, W. (2017). Position-aware activity recognition with wearable devices. *Pervasive and Mobile Computing*, *38*, 281–295. https://doi.org/10.1016/j.pmcj.2017.01.008

[10] Oyeleye, M., Chen, T., Titarenko, S., & Antoniou, G. (2022). A predictive analysis of heart rates using machine learning techniques. *International Journal of Environmental Research and Public Health*, *19*(4), 2417.

[11] Manjarres, J., Narvaez, P., Gasser, K., Percybrooks, W., & Pardo, M. (2019). Physical workload tracking using human activity recognition with wearable devices. *Sensors*, *20*(1), 39. https://doi.org/10.3390/s20010039

[12] Choksatchawathi, T., Ponglertnapakorn, P., Ditthapron, A., Leelaarporn, P., Wisutthisen, T., Piriyajitakonkij, M., & Wilaiprasitporn, T. (2020). Improving heart rate estimation on consumer grade wrist-worn device using post-calibration approach. *IEEE Sensors Journal*, *20*(13), 7433–7446.

[13] Fuller, D., Anaraki, J. R., Simango, B., Rayner, M., Dorani, F., Bozorgi, A., …, & Basset, F. A. (2021). Predicting lying, sitting, walking and running using Apple Watch and Fitbit data. *BMJ Open Sport & Exercise Medicine*, *7*(1), e001004.

[14] Husnain, A., Hussain, H. K., Shahroz, H. M., Ali, M., & Hayat, Y. (2024). A precision health initiative for chronic conditions: Design and cohort study utilizing wearable technology, machine learning, and deep learning. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(2), 118–139.

[15] Damre, S. S., Shendkar, B. D., Kulkarni, N., Chandre, P. R., & Deshmukh, S. (2024). Smart healthcare wearable device for early disease detection using machine learning. *International Journal of Intelligent Systems and Applications in Engineering*, *12*, 158–166.

[16] Prabhavathi, K., Selvi, K., Poornima, K. N., & Sarvanan, A. (2014). Role of biological sex in normal cardiac function and in its disease outcome–a review. *Journal of Clinical and Diagnostic Research*, *8*(8), BE01.

[17] Schäppi, J., Stringhini, S., Guessous, I., Staub, K., & Matthes, K. L. (2022). Body height in adult women and men in a cross-sectional population-based survey in Geneva: Temporal trends, association with general health status and height loss after age 50. *BMJ Open*, *12*(7), e059568. https://doi.org/10.1136/bmjopen-2021-059568

[18] Ong, Y. J., Baracaldo, N., & Zhou, Y. (2022). Tree-based models for federated learning systems. In H. Ludwig & N. Baracaldo (Eds.), *Federated learning: A comprehensive overview of methods and applications*. Springer International Publishing.

[19] Kurama, V. (2019). A guide to random forests: Consolidating decision trees. *Paperspace Blog*. Retrieved from: https://blog.paperspace.com/random-forests

[20] Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). CHIRPS: Explaining random forest classification. *Artificial Intelligence Review*, *53*, 5747–5788.

[21] Borup, D., Christensen, B. J., Mühlbach, N. S., & Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, *39*(2), 841–868.

[22] Khoei, T. T., Ismail, S., & Kaabouch, N. (2021). Boosting-based models with tree-structured parzen estimator optimization to detect intrusion attacks on smart grid. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, 0165–0170.

[23] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967.

[24] Khandelwal, N. (2020). *A brief introduction to XGBoost. Extreme gradient boosting with XGBoost! | by Neetika Khandelwal. Towards data science*. Retrieved from: https://towardsdatascience.com/a-brief-introduction-to-xgboost-3eaee2e3e5d6

[25] Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, *115*(532), 1978–1997.

[26] Parvandeh, S., Yeh, H. W., Paulus, M. P., & McKinney, B. A. (2020). Consensus features nested cross-validation. *Bioinformatics*, *36*(10), 3093–3098.

How to Cite: Yusuf, A., Al Jaber, T., & Gordon, N. (2025). Comprehensive Health Tracking Through Machine Learning and Wearable Technology. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS52023588