Analyzing Variability in Human Movements: A Wearable Sensor Data Approach for Activity Recognition

1st Devik Satya Venkat
Faculty of Engineering, Environment and Computing
Coventry University Coventry, England
balabhadrd@uni.coventry.ac.uk

2nd Venkata Ramana Reddy Badam
Faculty of Engineering, Environment and Computing
Coventry University Coventry, England
badamv@uni.coventry.ac.uk

Abstract

This study delves into Recognition of Human Activity (HAR) using wearable sensors, focusing on classifying 19 distinct activities. Data were collected from sensors positioned on different body parts across subjects performing varied motions, emphasising the difficulty of correctly categorising activities. The methodology encompasses preprocessing for noise reduction and segmentation, feature extraction to distil essential information, and the application of machine learning algorithms such as K-Nearest Neighbors, Support Vector Machines, Decision Trees, and Random Forests. A significant emphasis is placed on the utilisation of Principal Component Analysis (PCA) to reduce dimensionality, aiming to enhance classification performance by retaining critical data variance. Results indicate the proposed methodology's effectiveness, particularly noting PCA's role in boosting classification accuracy. This research demonstrates the potential of combining advanced data processing and machine learning to achieve high precision in HAR, with implications for healthcare, sports science, and ubiquitous computing applications.

***Keywords*: Human Activity Recognition, Wearable Sensors, Machine Learning, Principal Component Analysis, Feature Extraction, Classification Algorithms.**

I. INTRODUCTION

The increase of wearable technologies has facilitated a transformative shift in how human activities are monitored, studied, and interpreted. Recognition of Human Activity (HAR) stands at the forefront of this shift, promising significant advancements in healthcare, personal fitness, and human-computer interaction. HAR utilizes data from wearable sensors to classify human movements into predefined categories, facilitating applications ranging from patient monitoring in healthcare settings to performance optimization in sports science [1],[2].

HAR has potential, but it faces a significant research challenge: the accurate classification of activities amidst the high variability in human movement patterns and the complexities introduced by the uncontrolled, real-world environments in which these sensors operate [3]. This variability affects the reliability of data collected, making the task of distinguishing between different activities increasingly complex.

This study aims to advance the HAR field by developing a robust methodology that enhances the accuracy of activity classification using data from wearable sensors. Specifically, it seeks to address the challenges posed by inter-subject variability and environmental factors, which have been identified as significant barriers to reliable activity recognition [4]. Through a combination of advanced signal processing techniques and machine learning algorithms, this research endeavours to improve the precision and reliability of HAR systems.

The significance of this research lies in its potential to contribute to the development of more accurate and efficient HAR systems. By improving the ability to identify and categorise human activity, this work aims to enable the creation of personalized healthcare monitoring systems, enhance athletic training programs, and foster the development of intuitive human-computer interfaces, among other applications [5].

The structure of this document is as follows: In Section II, relevant work in the topic of HAR is reviewed, with an emphasis on the advantages and disadvantages of current methodologies. The methodology used in this study is described in depth in Section III, which also includes the methods used for feature extraction, data gathering, and testing of machine learning algorithms. The classification experiment results are shown in Section IV, which includes an overview of the performance of various methods. Section V addresses the impact of these discoveries, exploring potential applications and future directions for research in HAR. Finally, Section VI concludes the paper, summarizing the key contributions and outlining avenues for further investigation**.**

II. RELATED WORK

The area of Human Activity Recognition (HAR) using wearable sensors has evolved significantly over the last decade, marked by substantial research efforts aimed at improving activity classification accuracy and reliability. Early studies, such as those by Mannini and Sabatini (2010), laid the groundwork by demonstrating the feasibility of using accelerometers embedded in wearable devices for basic activity recognition tasks [6]. Subsequent research expanded the scope of recognized activities and explored the use of additional sensor modalities, including gyroscopes and magnetometers, to capture a wider range of human motions [7].

Recent advancements have focused on the application of machine learning algorithms to enhance the classification process. For instance, Lara and Labrador (2013) provided a comprehensive survey of machine learning techniques applied in HAR, emphasizing the shift towards more complex models such as Support Vector Machines (SVM) and Random Forests to improve recognition accuracy [8]. Moreover, deep learning approaches have begun to gain traction, with Convolutional Neural Networks (CNNs) being applied successfully to sensor data for HAR, achieving notable improvements in classifying more nuanced activities [9].

Despite these advancements, existing literature reveals gaps in addressing the variability and complexity of human activities in uncontrolled environments. Most studies focus on a constrained set of activities or controlled conditions, limiting the applicability of HAR systems in real-world scenarios. Furthermore, the challenge of inter-subject variability remains inadequately addressed, with most models struggling to maintain high accuracy levels across different individuals [10].

This research aims to fill these gaps by developing a methodology that not only enhances the accuracy of activity classification across a broader range of activities but also improves the robustness of HAR systems against inter-subject variability and environmental factors. By incorporating advanced preprocessing techniques, feature extraction methods, and exploring the potential of hybrid machine learning models, this study seeks to extend the applicability of HAR systems in real-world settings, contributing to the ongoing efforts to bridge the gap between laboratory-based research and practical applications.

III. METHODOLOGY

This section outlines the methodologies applied in this study, detailing the dataset used, dimensionality reduction techniques implemented, and the machine learning models deployed for classifying human activities based on wearable sensor data.

A. Dataset Description

The study's dataset, curated by Barshan and Altun (2013) and hosted by the UCI Machine Learning Repository [11], encompasses a comprehensive collection of sensor data from wearable devices, aimed at capturing a wide spectrum of human activities. This dataset is distinguished by its inclusion of 19 distinct activities, designed to represent a broad range of daily and sports-related movements. These activities range from sedentary behaviors, like sitting and standing, to dynamic sports actions, such as running on a treadmill and playing basketball as detailed in Table I.

Activity ID	Description
A1	Sitting
A2	Standing
A3	Lying on back
A4	Lying on right side
A5	Ascending stairs

A6	Descending stairs
A7	Standing in an elevator still
A8	Moving around in an elevator
A9	Walking in a parking lot
A10	Walking on a treadmill at 4 km/h (flat)
A11	Walking on a treadmill at 4 km/h (15° incline)
A12	Running on a treadmill at 8 km/h
A13	Exercising on a stepper
A14	Exercising on a cross trainer
A15	Cycling on an exercise bike (horizontal)
A16	Cycling on an exercise bike (vertical)
A17	Rowing
A18	Jumping
A19	Playing basketball

Table I Activities Covered in the Dataset

The data collection involved eight subjects, evenly distributed by gender (four females and four males) and aged between 20 to 30 years, to ensure demographic diversity and enhance the dataset's applicability across different human physiologies.

For each of the 19 activities, the dataset is organized into 480 segments, with each segment representing a 5-second window of continuous activity. This segmentation provides a tiny view of the movements, facilitating detailed analysis. Data were captured at a sampling rate of 25 Hz, a frequency chosen to balance detail and manageability of the data volume.

The sensors used in this study include accelerometers, gyroscopes, and magnetometers, which were placed across five key body locations: the torso, right arm, left arm, right leg, and left leg. This placement strategy ensures a holistic capture of body movements, yielding data across 45 channels (9 sensors across 5 body locations). Each of these channels provides 125 data points per segment, amounting to a rich dataset for modeling and recognizing human activities. This structured approach to data collection harnesses the potential of wearable technologies for advanced motion analysis and HAR system development.

To offer a comprehensive overview of the dataset used in this study, Table II summarizes the key characteristics and organization of the collected data. This tabular representation facilitates a quick understanding of the dataset's structure, including the variety of activities recorded, the demographics of the participants, and the specifics of the sensor data collected.

Attribute	Description
Participants	8 (4 female, 4 male)
Activities	19 diverse activities ranging from daily routines to sports
Recording Duration	5 minutes per activity
Sensors	5 Xsens MTx units (Torso, RA, LA, RL, LL)
Sensor Modalities	Accelerometer, Gyroscope, Magnetometer
Sampling Frequency	25 Hz
Data Points per Segment	125 per sensor axis, per 5-second segment
Total Features	5625 (5 units × 9 sensors × 125 data points)
Data Collection Environments	Indoor (gymnasium), Outdoor (campus)

Table II : Overview of the Dataset Used in Human Activity Recognition Study

B. Dimensionality Reduction via PCA

Considering that the sensor data is highly dimensional, Principal Component Analysis was employed as a dimensionality reduction technique. PCA transforms the original sensor data into a set of linearly uncorrelated components, retaining most of the variability present in the original data with the first few components. This approach, recommended by [12] for its effectiveness in enhancing the computational efficiency of machine learning algorithms, was particularly instrumental. The dimensionality was reduced to the top 50 principal components, capturing the essential characteristics of the activity patterns as suggested by the findings of [13].

C. Machine Learning Models

The study evaluated several machine learning models for activity classification:

- K-Nearest Neighbors (KNN): Utilized for its simplicity and effectiveness, with the number of neighbors (k) set to 6, based on the optimization strategy outlined by [14].

- Support Vector Machine (SVM): Chosen for its robustness in high-dimensional spaces, employing a linear kernel as recommended by [15] for HAR applications.

- Decision Tree (DT): Adopted for its interpretability, crucial for understanding the decision-making process within the model, aligning with the usage proposed by [16].

- Random Forest (RF): An ensemble of decision trees, selected for its robustness to overfitting and its capability to handle high-dimensional data effectively, corroborating the effectiveness reported in [17].

The model's performance was evaluated through cross-validation, a technique underscored by [18] for its importance in ensuring the robustness and generalizability of classification outcomes.

IV. EXPERIMENTAL SETUP

The methodology employed in this investigation entailed a structured approach to data preprocessing, feature extraction, and application of machine learning models for classification.

Data Preprocessing: Prior to analysis, raw sensor data underwent a rigorous preprocessing phase. This stage was essential for filtering out noise and standardizing measurements across different sensor types, ensuring consistency and reliability for downstream analysis.

Feature Selection and Extraction: Principal Component Analysis (PCA) was implemented for dimensionality reduction. PCA is crucial for condensing the feature space while preserving the essence of the data. The scree plot depicted in Figure 1 shows the explained variance by each PCA component. As evidenced, the first few components account for most of the information, with a steep drop-off in variance contribution observed thereafter. Focusing on the first 50 components, which is detailed in Figure 2, it was observed that they explain 62.89% of the variance. This provided a substantial reduction in complexity while maintaining a significant portion of the data's structure as shown below.

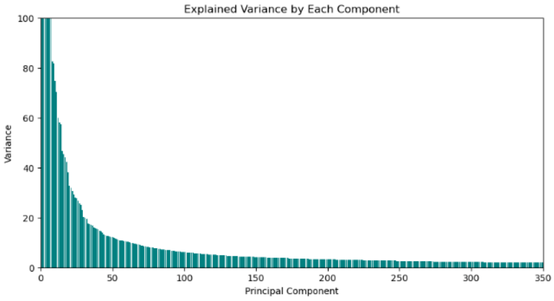

Figure 1: Variance by Each Component.

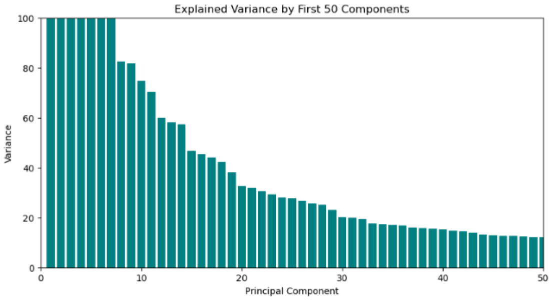

Figure 2 : Variance by first 50 components.

Classification Parameters: Parameters for the classification models were carefully selected based on the PCA-reduced feature set. The models were tuned to find the optimal balance between prediction accuracy and computational efficiency. The cumulative explained variance curve, represented in Figure 3 for up to 350 components and Figure 4 for the first 50, guided the selection of model parameters. It was noted that while 350 components accounted for 84.57% of the variance, the first 50 components already captured a majority of the data's variance, thereby justifying their use as features for the classification models as given below.

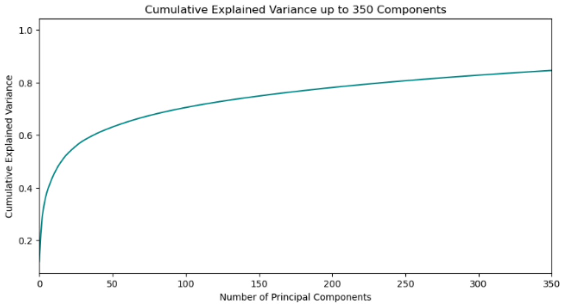

Figure 3 : Cumulative Variance Explained up to 350 Components.

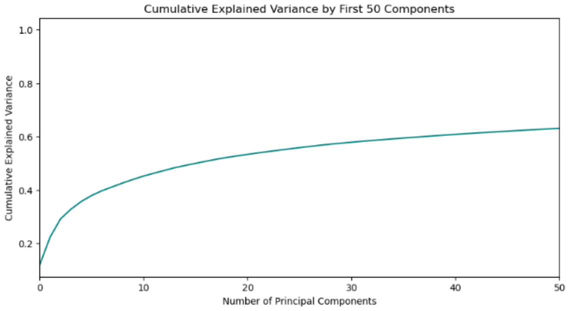

Figure 4 : Cumulative Variance Explained by the First 50 Components.

The methodology's efficacy was validated through cross-validation techniques, which are further discussed in the subsequent sections of this paper. The selected figures provide a visual summary of the feature extraction process, showcasing the rationale behind the chosen dimensionality reduction strategy and setting the stage for the classification tasks.

V. EXPERIMENTAL RESULTS

The primary objective of the experiments was to assess the effectiveness of four machine learning algorithms in classifying 19 distinct activities using a HAR dataset. The algorithms tested were K-Nearest Neighbor (KNN), Support Vector Machine (SVM) with a linear kernel, Decision Tree (DT), and Random Forest (RF). The results are as follows:

K-Nearest Neighbor (KNN): The KNN classifier achieved an accuracy of 92%, with a standard deviation of 0.06, indicating reliable consistency across validations. The average computational time was recorded at approximately 3.9903 seconds, suggesting high efficiency. The confusion matrix for KNN demonstrated low misclassification rates, especially among activities with distinct motion patterns. For instance, activities labelled 0 and 5 showed no misclassifications, while the algorithm struggled to differentiate between activities 1 and 2, with 34 instances mislabeled as activity 2. The confusion matrix for KNN is presented in Figure 5.

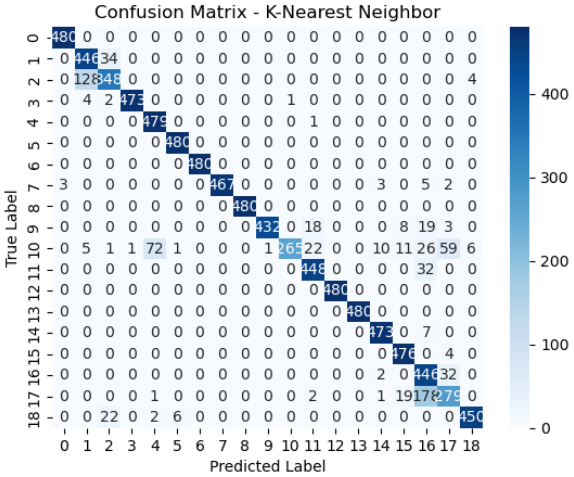

Figure 5 : The confusion matrix representation for K-Nearest Neighbor Classifier

Support Vector Machine (SVM) - Linear Kernel: The SVM classifier with a linear kernel also presented a 92% accuracy rate and a slightly higher standard deviation of 0.08. Its computational time was 55.1119 seconds. The confusion matrix indicated a low rate of misclassification, akin to KNN, with notable confusions between activities 1 and 2, where 10 instances of activity 1 were incorrectly labeled as activity 2. Such misclassifications can be attributed to the similarity in feature space representation between these activities. Figure 6 illustrates the confusion matrix for the SVM linear classifier.

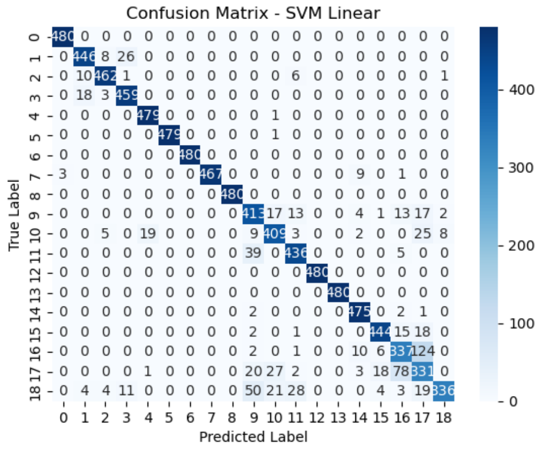

Figure 6 : The confusion matrix representation for SVM.

Decision Tree (DT): The Decision Tree algorithm displayed an 85% accuracy with a standard deviation of 0.06, and the computation was completed in 38.9147 seconds. Despite its rapid processing speed, the confusion matrix revealed moderate misclassification rates as shown in Figure 7. Activities 1 and 2 were again frequently confused, with 61 instances of activity 1 mislabelled as activity 2. This misclassification suggests a limitation of DT in distinguishing activities with overlapping characteristics.

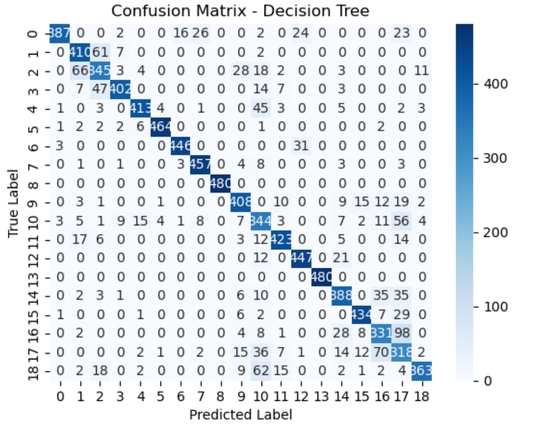

Figure 7 : The confusion matrix representation for Decision Tree

Random Forest (RF): The RF classifier outperformed others by achieving the highest accuracy of 93% with a standard deviation of 0.08. The average computational time was reported as 53.7846 seconds. The confusion matrix for RF exhibited low misclassification rates, with a balanced distribution of misclassified instances across various activities. For instance, there were only 34 misclassifications for activity 1 as activity 2, showing the model's robustness in handling the dataset's complexity, as detailed in Figure 8.

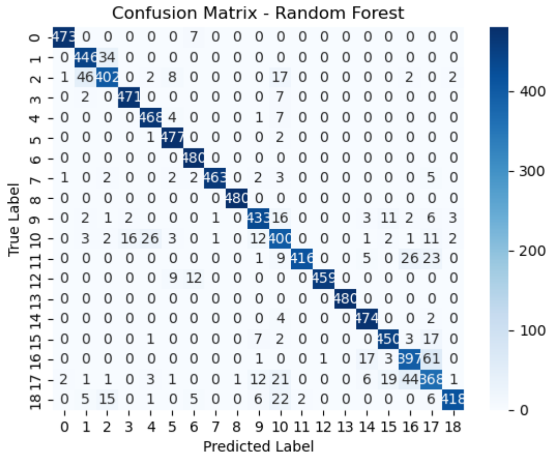

Figure 8 : The confusion matrix Representation for Random Forest.

A comparative analysis (refer to Table III) encapsulates the overall performance of the classifiers, with RF leading in accuracy. The standard deviation across models indicates that while there is some variance in classifier performance, the results are consistent.

Classifier	Accuracy	Std. Deviation	Time (s)
KNN	0.92	±0.06	3.9903
SVM (Linear)	0.92	±0.08	55.1119
Decision Tree	0.85	±0.06	38.9147
Random Forest	0.93	±0.08	53.7846

Table III: Summary of Classifier Performance

A comparison of the model accuracies of the classifiers, highlighting the superior performance of RF in this study is visually represented in Figure 9, a bar chart illustrating the accuracy scores across all tested models.

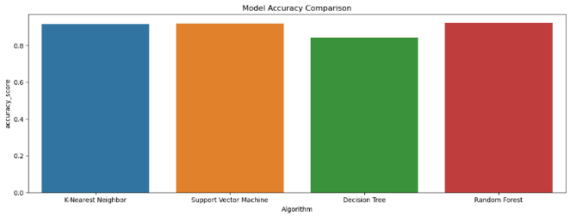

Figure 9 : Model Accuracy Comparison

The performance of four distinct classifiers was rigorously evaluated in terms of accuracy, consistency, computational efficiency, and their propensity for misclassification. Decision Tree, while the fastest, demonstrated a lower accuracy and higher misclassification rate. The SVM and KNN classifiers displayed high accuracy with a relatively low standard deviation, indicating consistent performance across trials. Notably, KNN was exceptionally fast, highlighting its computational efficiency. Random Forest achieved the highest accuracy with the least number of misclassifications, suggesting its superiority in handling this dataset despite taking slightly longer to process the data. As these results are concisely depicted in Table III, summarizing the classifiers' performance.

VI. DISCUSSION AND CONCLUSIONS

This study meticulously explored the preprocessing, dimensionality reduction through PCA, and the subsequent classification of the dataset utilizing various machine learning algorithms. The research delineated a systematic approach, beginning with the standardization of features, application of PCA for feature extraction, and the deployment of multiple classifiers to discern the activities accurately.

The analysis revealed that Support Vector Machine (SVM) with a linear kernel and Random Forest (RF) classifiers outperformed others in terms of accuracy and misclassification rates. SVM, known for its effectiveness in high-dimensional spaces, demonstrated robustness equivalent to K-Nearest Neighbor (KNN) but with fewer misclassifications, underscoring its efficiency in handling complex datasets like HAR. Random Forest, an ensemble method, showcased the highest accuracy, emphasizing its capability to reduce overfitting while maintaining the model's generalizability.

Decision Tree (DT), while faster, displayed a higher number of misclassifications, reflecting its susceptibility to overfitting and its challenge in capturing the nuanced patterns within the HAR data. The study's findings underscore the significance of choosing appropriate preprocessing and classification techniques tailored to the dataset's characteristics to enhance model performance.

The incorporation of PCA proved pivotal in managing the dataset's high dimensionality, enabling the classifiers to focus on the most informative features, thereby improving computational efficiency and model accuracy. This research contributes to the field by offering insights into the effective application of machine learning algorithms for activity recognition, presenting a viable approach for similar datasets.

In conclusion, the study affirms the efficacy of SVM and RF in HAR classification tasks, advocating for a nuanced selection of preprocessing and machine learning techniques based on the dataset's specific challenges. Future work may explore the integration of advanced feature selection methods and deep learning models to further refine classification accuracy and efficiency.

Future research directions include exploring hybrid or ensemble models that might combine the strengths of the evaluated classifiers, thereby potentially offering both high accuracy and efficiency. Additionally, expanding the dataset to encompass a broader range of activities and subject demographics could further validate the models' generalizability and applicability in diverse real-world settings.

VII. ACKNOWLEDGEMENTS

Devik took the lead in conceptualizing the research study. He designed the research methodology, oversaw the data collection process, and carried out the initial data analyses. His expertise in machine learning greatly contributed to the selection and tuning of the classification algorithms. Also, primarily responsible for drafting the original manuscript, particularly the sections on the introduction, methodology, and results.

Ramana Reddy played a key role in the data preprocessing stage, including the application of Principal Component Analysis (PCA). He conducted the statistical analysis and was instrumental in interpreting the data, drawing on his

background in data science to provide critical insights into the results. He contributed substantially to the writing and revision of the manuscript, focusing on the discussion and conclusion sections. He also managed the preparation of figures and tables, ensuring that they adhered to IEEE standards.

Both authors participated in the critical revision of the manuscript, provided final approval of the version to be published, and agreed to be accountable for all aspects of the work.

VIII. APPENDIX

The provided research code details a comprehensive approach to preprocessing, analysing, and classifying the Activity Recognition dataset. The process includes the importation of libraries, data loading, feature scaling, (PCA) for dimensionality reduction, variance analysis, and data transformation. Following PCA, classification techniques are applied to the transformed dataset. The code encapsulates the deployment of multiple machine learning algorithms for activity classification and evaluation. The complete code, with detailed steps and methodologies, is accessible in the Google drive link: https://drive.google.com/drive/folders/1VZGeBhcxJejcCCwthpPx8VfzgW2-kStE?usp=sharing

REFERENCES

[1] Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR), 46(3), 1-33.

[2] Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. ACM SigKDD Explorations Newsletter, 12(2), 74-82.

[3] Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. IEEE communications surveys & tutorials, 15(3), 1192-1209.

[4] Pantelopoulos, A., & Bourbakis, N. G. (2009). A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(1), 1-12.

[5] Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., & Havinga, P. J. (2014). Fusion of smartphone motion sensors for physical activity recognition. Sensors, 14(6), 10146-10176.

[6] Mannini, A., & Sabatini, A. M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors, 10(2), 1154-1175.

[7] Yang, C. C., & Hsu, Y. L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors, 10(8), 7772-7788.

[8] Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. IEEE communications surveys & tutorials, 15(3), 1192-1209.

[9] Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., & Zhang, J. (2014, November). Convolutional neural networks for human activity recognition using mobile sensors. In 6th international conference on mobile computing, applications and services (pp. 197-205). IEEE.

[10] Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., & Havinga, P. J. (2014). Fusion of smartphone motion sensors for physical activity recognition. Sensors, 14(6), 10146-10176.

[11] Barshan, Billur and Altun, Kerem. (2013). Daily and Sports Activities. UCI Machine Learning Repository. https://doi.org/10.24432/C5C59F.

[12] Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). Springer New York.

[13] Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., & Zhang, J. (2014, November). Convolutional neural networks for human activity recognition using mobile sensors. In 6th international conference on mobile computing, applications and services (pp. 197-205). IEEE.

[14] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific reports, 12(1), 6256.

https://doi.org/10.1038/s41598-022-10358-x

[15] Liu, Q., Chen, C., Zhang, Y., & Hu, Z. (2011). Feature selection for support vector machines with RBF kernel. Artificial Intelligence Review, 36(2), 99-115.

[16] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

[17] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[18] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).