Distance-based wrapper model for human activity recognition

Rana Alauldeen Abdalrahman [a,1,*], Laith AL-Frady [b,2], Ruaa Ali Khamees [c,3]

[a] Middle Technical University, ICT Department, Polytechnic College of Engineering-Baghdad, Iraq
[b] IT Department, Technical College of Management, Middle Technical University, Baghdad, Iraq
[c] ICT Department, Middle Technical University, Baghdad, Iraq
[1] rana.ala@mtu.edu.iq; [2] al_frady_laith@mtu.edu.iq; [3] dr.ruaakhamees@mtu.edu.iq
* corresponding author

ARTICLE INFO

Article history
Received January 6, 2026
Revised February 13, 2026
Accepted March 17, 2026
Available online May 31, 2026

Keywords
Human activity recognition
Smartphone inertial sensors
Feature selection
Classification

ABSTRACT

In this paper, we developed an effective wrapper-based model to optimize the recognition of physical human daily life activities recorded by smartphone built-in sensors. The proposed model employs the Sequential Forward Selection (SFS) method in combination with the K-Nearest Neighbor (KNN) classifier based on a group of five commonly used distance measures: (Manhattan, Euclidean, Chebyshev, Canberra, and Correlation), each having a specific geometric neighborhood interpretation. The proposed SFS-KNN multi-distance approach allows each distance metric to guide feature selection. It examines how the resulting feature subsets, determined by each distance metric, affect overall recognition performance. The goal is to identify the best feature subset that achieves the highest accuracy with the lowest dimensionality. Our proposed distance-based wrapper model was validated on two publicly available WISDM and UCI-HAR datasets under 10-fold cross-validation. The experimental results on both datasets showed that the proposed model's performance is significantly affected by the choice of distance measure, as different feature subsets are generated for each distance measure. For the WISDM dataset, our model achieves an overall accuracy of 93.39% based on Euclidean distance, with a reduction ratio of 71.42%. It also achieves a 95.9% reduction in feature dimensionality on the UCI-HAR dataset using the Correlation distance, with a recognition rate of 99.33%. These outcomes confirm the superiority of our wrapper model over other feature selection-based approaches proposed for the same datasets.

1. Introduction

The recognition of the daily activities routinely performed by humans, often referred to as Activities of Daily Living (ADLs), has attracted considerable attention in recent years. Thus, a new gateway to a fast-growing research field, Human Activity Recognition (HAR), has been opened. This active field presents promising potential to enhance the quality of human life via a broad range of application areas such as healthcare monitoring, fall detection of elderly people, fitness tracking, smart environments, and homeland security [1], [2]. These areas highlight the importance of HAR models as assistive tools for understanding human behavior and actions.

The ADL recognition process commonly starts by recording body motion signals from wearable sensors, ambient sensors, or both [3], [4], [5], [6]. First, the recorded signals are partitioned into equal segments using a fixed-size sliding window. Then, the obtained raw segments are analyzed to extract appropriate features via various feature extraction methods. Finally, the extracted features are used as

inputs to machine learning approaches to yield the ultimate HAR model. The segment length and the quality of the generated features play a vital role in determining the overall model performance [7]. Ambient sensors often rely on video cameras as a convenient means to capture human physical activities in the ambient environment [8]. However, the disturbance they cause to human privacy due to their intrusiveness has limited their ubiquitous use. On the contrary, considerable attention has been paid to the use of various wearable sensor-based technologies, such as inertial, vital-signs, and GPS sensors, as less invasive and more acceptable alternatives for HAR [8], [9].

Recently, various inertial sensors, to name but a few, accelerometers, gyroscopes, and magnetometers, have been either integrated into ad-hoc sensors (e.g., IEE smart insoles and Shimmer) or built into modern mobile devices (e.g., smartphones and smartwatches) [10], [11]. These devices have been successfully used in real-world scenarios, alone or in combination, to differentiate among multiple basic activities, which can be static (e.g., sitting, standing, and lying) or dynamic (e.g., walking, running, and stair climbing). Compared with ad-hoc devices, the acceptability, ease of use, size, cost, processing, and wireless capabilities are the main practical and technical aspects that have made smartphones the best candidate sensing devices for unobtrusive activity recognition [12], [13]. More importantly, previous studies demonstrated that body motion data retrieved by smartphone inertial sensors are adequately accurate for HAR [14], [15].

Despite the aforesaid advantages, the availability of data collected from smartphone-based sensors is relatively limited. This is because the majority of HAR literature has focused on vision-based sensors for data acquisition [13]. Another significant reason is that most researchers conduct smartphone-based HAR experiments based on their proprietary datasets, which are rarely available to the public domain. As a result, this prevents the research community interested in this field from making objective comparisons between the recently proposed HAR approaches and the previous ones. To facilitate benchmarking studies, however, some research papers have offered free access to standard datasets: WISDM [16], UCI-HAR [17], and HARTH [18], containing labeled inertial sensors data that were recorded from groups of subjects performing a specific set of ADLs while carrying their smartphones.

Although it has been well-addressed in literature, the task of HAR using smartphones' inertial sensors is still of great interest to many researchers. Previous publications have shown the capability of several classification models in offering successful recognition results for this task [13], [19], [20]. Recently, more attention has been paid to investigating the feasibility of utilizing various feature selection methods (filter-based, wrapper-based, or a combination of both) with ML and DL approaches to improve the recognition of a set of daily human activities [21], [22]. This section presents a detailed revision of all previously proposed feature selection-based models for WISDM and UCI-HAR datasets on basis of the best optimization method used, the optimal features obtained, the validation criterion adopted, and the highest recognition accuracy achieved.

For WISDM dataset, the authors in [23] achieved 93.12% recognition accuracy with 11 features only using a combination of Sequential Forward Selection (SFS) and Gradient Boosted trees (GBT). The proposed SFS-GBT model was evaluated under 10-fold cross-validation criterion. Significant improvement of more than 98% overall accuracy was obtained on the same dataset in [24] and [25], but applying a smaller time window of size 2.56s with 50% overlapping for data segmentation than the standard 10s adopted in the original paper [16]. Therefore, direct comparison with the recognition results of these two studies cannot be conducted due to generating feature vectors with larger dimensions.

For UCI-HAR dataset, the authors in [26] combined Particle Swarm Optimization (PSO) method with K-Nearest Neighbors (KNN) based on Manhattan distance and achieved an accuracy rate of 98.89% with 239 features under 10-fold cross-validation. In another study on the same dataset [27], a Random Forest (RF) based feature selection method called Guided Regularized Random Forest (GRRF) was proposed to identify the most relevant features for activity recognition and eliminate those with minimum importance. For classification, Support Vector Machine (SVM) produced the best-reported

accuracy of 99.3% using 185 features only. The 70 % training and 30% testing validation scheme was adopted to evaluate the effectiveness of the proposed model.

Deep learning approaches namely, Recurrent Neural Network (RNN) and Deep Q-network (DQN) have also been integrated with feature selection algorithms to improve the classification performance of the set of activities recorded in WISDM and UCI-HAR datasets [28], [29]. In [28], the authors proposed a new metaheuristic method called Colliding Bodies Optimization (CBO) and achieved mediocre recognition results of about 90% on both datasets. Moreover, the authors did not report the evaluation strategy and the final feature subset used to produce these accuracy scores. In [29], Bee Swarm Optimization (BSO) was proposed to reduce UCI-HAR features using 70 % training and 30% testing split ratio and an average accuracy of 98.41% was achieved. However, the authors did not mention the reduced feature set used to generate this accuracy. More details of the optimization results obtained in previous studies on the two datasets are presented in Table 1.

Table 1. Previous feature selection methods for WISDM and UCI-HAR datasets.

Paper	Dataset	Proposed Method	Optimal Features	Testing Approach	Accuracy (%)
[23]	WISDM	SFS+GBT	11	10-fold cross-validation	93.12
[24]	UCI-HAR	GB OGWO+SVM	304	70% training, 30% testing	98
[25]	UCI-HAR	WOA+RF	236	70% training, 30% testing	95.51
[26]	WISDM	PSO+KNN	8	10-fold cross-validation	91.07
	UCI-HAR		239		98.89
[27]	UCI-HAR	GRRF+SVM	185	70% training, 30% testing	99.30
[28]	WISDM	CBO+RNN	Not listed	Not explained	89.77
	UCI-HAR				90.19
[29]	UCI-HAR	BSO+DQN	Not listed	70% training	98.41
[30]	UCI-HAR	FDDRT+RF	66	70% training, 30% testing	98.72
[31]	UCI-HAR	CGA+MLP	386	70% training, 30% testing	95.79
[32]	WISDM	BBO+RF	Not listed	80% training, 20% testing	79.38

Despite significant efforts in feature selection methods for human activity recognition using datasets like WISDM and UCI-HAR, several critical challenges remain unresolved. Most prior studies did not specify the number of optimal features obtained or used a custom split of the data, making it difficult or impossible to reproduce and compare their results. While some studies achieved high classification accuracy, this came at the cost of large feature sizes or training on a small portion of the entire data. Hence, compromising the generalization capability of their methods. These limitations call for an effective feature selection approach that can reduce dimensionality while maintaining or enhancing classification performance. To address such challenges, this paper proposes an efficient wrapper-based approach integrating Sequential Forward Selection (SFS) with K-Nearest Neighbors (KNN) to enhance the recognition accuracy of the set of ADLs recorded in the two popular datasets mentioned above. The proposed model adopts five different distance measures, namely Manhattan, Euclidean, Chebyshev, Canberra, and Correlation distances, to investigate the impact of the generated per-distance feature subset on the overall classification performance. The overall contributions of this paper are summarized as:

- A distance-conditioned wrapper feature selection study showing that the distance metric changes the selected subset, and this materially changes accuracy and dimensionality reduction.

- A fair benchmark protocol across two canonical HAR datasets under the same validation regime (10-fold CV) and the same wrapper search strategy, with feature subsets reported.
- A considerable reduction in feature dimensionality is attained by selecting the most discriminant feature subset per distance from the original feature spaces of both datasets.
- A substantial improvement in the recognition accuracy is achieved compared to the previous feature selection HAR models that were trained and evaluated on the same datasets.

The rest of this paper is organized as follows: Section 2 demonstrates a detailed description of the proposed methodology. Section 3 shows the experimental results of feature selection and classification obtained on both datasets and compares them with those of existing alternative approaches. Finally, Section 4 summarizes the conclusions and outlines the future direction.

2. Method

This section introduces a comprehensive description of the five steps of the proposed methodology through which the present research was accomplished. The first step shows a brief overview of the main characteristics of the WISDM and UCI-HAR datasets. The second step shows the preprocessing mechanisms performed on WISDM data prior to classification. The third and fourth steps provide a detailed explanation of the distance-based wrapper model that combines the sequential forward selection method with k-nearest neighbor classifier using five distance measures: Manhattan, Euclidean, Chebyshev, Canberra, and Correlation distances, respectively. The fifth step demonstrates the evaluation metrics and the validation technique adopted for evaluating the performance of our wrapper model. Fig. 1 illustrates the proposed methodology followed in this paper. It is worth mentioning that the adoption of these five different distance functions, shown in Fig. 1, is to ensure a multi-perspective feature evaluation. This diversity is crucial for SFS-KNN, as it allows the model to evaluate feature subsets from different angles and identify the relevant features determined by the most appropriate distance metric for the data.

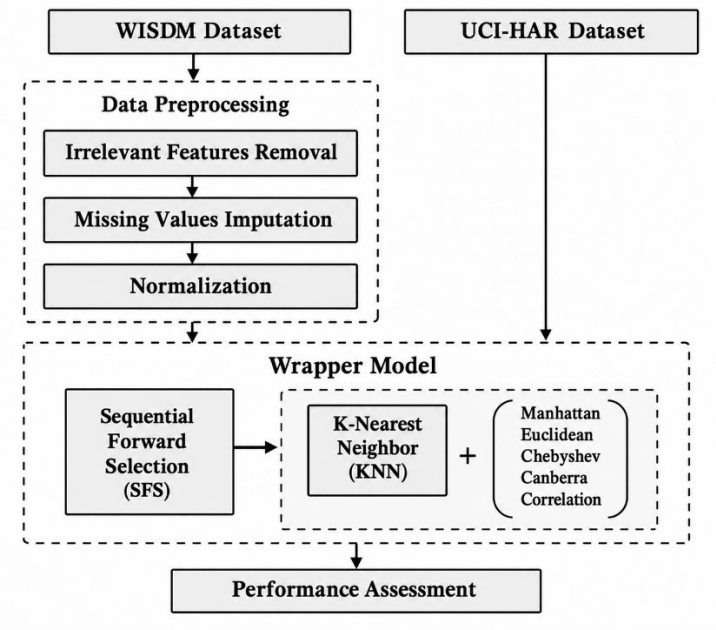

Fig. 1. Proposed methodology.

2.1. HAR Datasets Characteristics

In this paper, two publicly available Human Activity Recognition (HAR) datasets were employed to perform a set of feature selection and classification experiments. The first dataset, known as WISDM (Wireless Sensor Data Mining) [33], contains accelerometer-based data collected from thirty-six

volunteers, whereas the second dataset, UCI-HAR [34], incorporates gyroscope and accelerometer data gathered from thirty participants. The volunteers in both datasets collection sessions performed a certain group of physical daily life activities, often called Activities of Daily Living (ADL): (walking, jogging, laying down, stair climbing, sitting and standing) for a specific time while having their Android smartphones carried in their front pants pockets (for the WISDM data registration session) or mounted to their waists (for the UCI-HAR data recording session). Table 2 presents an overview of the main characteristics of the two datasets.

Table 2. Characteristics of WISDM and UCI-HAR datasets. [a] This type of activity was not recorded in the respective dataset.

Activity Type	WISDM Dataset	WISDM Dataset	UCI-HAR Dataset	UCI-HAR Dataset
# of features	**# of Instances**	**%**	**# of Instances**	**%**
	45		**561**	
Walking	2081	38.41	1722	16.72
Jogging	1625	29.99	–[a]	–[a]
Laying down	–[a]	–[a]	1944	18.88
Upstairs	632	11.66	1544	14.99
Downstairs	528	9.75	1406	13.65
Sitting	306	5.65	1777	17.25
Standing	246	4.54	1906	18.51
Total	**5418**		**10299**	

2.2. Preprocessing

Before conducting feature selection and classification experiments on the WISDM dataset, the following data preparation steps are performed. First, features such as UNIQUE_ID and USER contain only personal participant information, not accelerometer data. These two features are irrelevant to the classification analysis and should be removed from the dataset as a preliminary step. Another irrelevant feature is XAVG. This feature adds nothing to the discriminative power of the classification method, as it incorporates only zero values across all instances. Therefore, XAVG has also been eliminated. Secondly, three consecutive features namely, XPEAK, YPEAK, and ZPEAK are found to have notable missing values in various instances. Table 3 presents the total missing-valued instances per activity type.

Table 3. Missing-valued instances per class.

Activity	# of Instances	%
Walking	52	2.5
Jogging	13	0.8
Upstairs	45	7.12
Downstairs	21	3.98
Sitting	172	56.21
Standing	171	69.51
Total	**474**	**8.75**

Table 3 indicates that missing-valued instances account for 8.75% of the WISDM dataset (474 instances out of 5418). It can also be seen from the underlined values that most of these instances belong to the two non-dynamic activities (Sitting and Standing), despite having considerably fewer instances than the other four dynamic ones, as shown in Table 2. Accordingly, the missing values associated with the aforementioned three features are replaced by the estimated mean value of each feature vector independently. Thirdly, the dataset has 12 features: YAVG, ZAVG, XPEAK, YPEAK, ZPEAK, XABSOLDEV, YABSOLDEV, ZABSOLDEV, XSTANDDEV, YSTANDDEV, ZSTANDDEV, and RESULTANT with larger value ranges than the other 30 features. Therefore, to prevent these features from dominating the distance computation while preserving the original distribution of the data, large-valued features are scaled to the range [0,1] using min-max normalization expressed by Equation (1).

$$x' = \frac{x - \min_x}{\max_x - \min_x} \quad (1)$$

where x and x' are the original and scaled values of feature X, $\min_x$ and $\max_x$ are its minimum and maximum values, respectively.

2.3. Wrapper Model

With very large feature spaces, feature selection becomes an important and desirable step for developing classification models due to the effect of irrelevant and redundant features on the computational complexity, interpretability, and prediction performance of the model [35], [36]. In general, a classification model trained on a smaller feature space is often more robust and reproducible than that built on the original larger feature space [37]. Feature selection methods can be mainly categorized into filter-based and wrapper-based methods. Filter-based methods often rely on statistical measures and can be employed without a modeling algorithm. Though computationally efficient, these methods completely ignore the effect of selected features on the prediction performance of the learning model [38], [39]. In contrast to filter-based methods, wrapper-based methods iteratively take into consideration the impact of adding or removing a particular feature on the model performance. Therefore, these methods are computationally more expensive than filter methods, but are able to identify the optimal feature subset yielding the highest performance for the specific classification model [40]. This paper proposes a wrapper model that uses Sequential Forward Selection (SFS) in conjunction with K-Nearest Neighbors (KNN) classifier. The proposed SFS-KNN method adopts five different distance measures namely, Manhattan, Euclidean, Chebyshev, Canberra, and Correlation to analyze the effect of a selected subset of features based on a particular distance on the performance of the model. The following subsections present a detailed explanation of these algorithms.

2.3.1. Sequential Forward Selection

Sequential Forward Selection (SFS) is a greedy search technique used to reduce the dimensionality of the original feature space by iteratively selecting the subset of features most relevant to a classification problem. The smaller feature subspace is guaranteed not only to improve the computational efficiency of the classification model at a later stage, but also to yield a more general model [41], [42]. This technique does not exhaustively search every possible feature subset, which is practically prohibitive due to the huge number of feature subset combinations that can be created, even for medium-sized feature sets. Consequently, SFS is computationally tractable as compared to the Exhaustive Feature Selection (ESF) algorithm, which requires evaluation of all candidate subsets to find the optimal one [43]. The standard SFS technique, described by the pseudocode in Algorithm 1, takes the entire n-dimensional feature space as input and returns an m-dimensional feature subspace, where m < n. To achieve this reduction in dimensionality, SFS initializes with an empty feature set $Y_0 = \emptyset$, $m = 0$ and iteratively adds one feature, x^+ at a time to Y_m that maximizes a user-specified performance optimization function $J(\cdot)$, for instance, the predictive accuracy of a classification model. The procedure of progressively incorporating features into larger subsets continues until no further optimization in performance is achieved with the addition of other features.

Algorithm 1: Pseudo-code of the SFS technique [41].

Input: $Y = \{y_1, y_2, \ldots, y_n\}$
Process:
Step 1: Start with empty set $Y_0 = \emptyset$, $m = 0$
Step 2: Select the best next feature

$$x^+ = \arg\max_{x \in Y - Y_m} J(Y_m + x)$$

Update feature set

$$Y_{m+1} = Y_m + x^+;\ m = m + 1$$

Go to Step 2.
Output: $Y_m = \{y_j, | , j, =, 12 \ldots m\ y_j \in Y\}$

2.3.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is an instance-based supervised machine learning algorithm widely used for classification and regression due to its simplicity, robustness, and reasonable accuracy. KNN has been successfully used for various pattern classification problems including object recognition [44], pattern recognition [45], ranking models [46], and text categorization [47]. It has also been used in other application areas such as Bioinformatics and medicine [44], [48]. In contrast to many classification algorithms, KNN does not explicitly train a model. Instead, it stores all training set samples and waits until producing a test sample. Therefore, KNN is often considered as a lazy learning method [49]. KNN is a non-parametric technique [50]. This implies that no assumptions about the underlying distribution of samples are required. As a result, it is an effective option for any classification problem that does not involve prior knowledge of the data samples' functional form. The traditional KNN algorithm, summarized by the pseudocode in Algorithm 2, works as follows. Given a training set D, a test sample $q = (x, c')$, and the value K (the number of nearest neighbors). The algorithm adopts a user-specified distance function or similarity measure to calculate the distance between q and every training sample, $p = (y, c) \in D$ to identify its nearest-neighbor list, D_q. In this notation, x denotes the feature vector of the test sample q and c' is its class label. Similarly, y is the feature vector of the training sample p and c is its class label. Once the D_q list is determined, the test sample is classified by the majority class (i.e. the most frequent class) of its K closest neighbors:

$$c' = \underset{v}{\operatorname{argmax}} \sum_{(y_i, c_i) \in D_q} I(v = c_i) \quad (2)$$

where v is a class label, c_i is te class of the i^{th} nearest training sample, and $\boldsymbol{I}(\cdot)$ is an indication function that returns 1 if its argument is satisfied and 0 if not. When $K = 1$, then the class of the nearest training neighbor is assigned to q.

Algorithm 2 Pseudo-code of KNN algorithm [49].

Input:
Training set D, Test sample $q = (x, \hat{c})$, K
Process:
Compute $d(x, y)$, the distance between q and every training sample, $p = (y, c) \in D$
Select $D_q \subseteq D$, the set of K nearest training samples to q.
Output: $\hat{c} = \arg\max_{v} \sum_{(y_i, c_i) \in D_q} I(v = c_i)$

The distance measure used by the KNN algorithm, and the pre-defined number of nearest neighbors (K), have a considerable influence on its final classification performance [51], [52]. The classical approach of optimizing the accuracy of KNN is often conducted experimentally. That is, various K values are tested for each distance function on the dataset under consideration, and the one that yields the highest overall accuracy for a given distance will be chosen to define the model. However, some previous benchmarking studies have attempted new approaches for augmenting the performance of KNN, either by proposing a non-convex distance measure [53] or by investigating other KNN variants that mitigate the need for determining the value of K in advance [54]. Despite the success and rationale of these methods, experimental results show that no single optimal distance or K value applies to all real-world datasets. This conclusion is consistent with the "No Free Lunch" theorem. In this paper, five distance measures, as shown in Table 4, are evaluated with the proposed SFS-KNN model to determine the most appropriate distance for the two human activity recognition datasets. Furthermore, the K-neighbors value is set to 1 intentionally during feature selection to ensure that any performance improvement achieved by the generated feature subset must be due to the distance measure used.

Table 4 illustrates the mathematical formulation of each distance function $d(x, y)$ to estimate the farness or dissimilarity between x and y, where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are the feature vectors of the testing and training samples having n numeric values. The distance functions given

by Equations (3-5) belong to the Minkowski distance (or L_p norm) family [55]. This power distance is a generalized measure and can be formulated as:

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}}, \qquad p > 0 \tag{8}$$

It becomes Manhattan (also called L_1 norm or City block distance) when p="1" , Euclidean (also known as L_2 norm or Ruler distance) when p="2" , Chebyshev (also referred to as Lagrange or Maximum value distance) when p= ∞. The Canberra distance represented by Equation (6) belongs to the L_1 norm family of measures that includes various modified versions of Manhattan distance [55]. The Correlation distance presented in Equation (7) is a normalized version of the Pearson distance measure that is calculated by subtracting the Pearson's correlation coefficient (PCC) between two vectors from oe [56]. It is worth mentioning that these distances, except for the Correlation distance, satisfy the four metric properties and considered as metric distances [55], [56].

Table 4. Distance functions formulas.

Distance	Formula	Range	
Manhattan	$d(x,y) = \sum_{i=1}^{n}\|x_i - y_i\|$	[0, + ∞]	(3)
Euclidean	$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	[0, + ∞]	(4)
Chebyshev	$d(x,y) = \max_i\|x_i - y_i\|$	[0, + ∞]	(5)
Canberra	$d(x,y) = \sum_{i=1}^{n}\frac{\|x_i - y_i\|}{\|x_i\| + \|y_i\|}$	[0, + n]	(6)
Correlation	$d(x,y) = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\right)$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$.	[0, 1]	(7)

Intuitively, the mathematical definitions provided in Table 4 explain how each distance measure uniquely quantifies the closeness between data points. As a result, this creates a variation in performance among these distances according to their different sensitivity to the inherent characteristics of the HAR data. The Euclidean distance is highly sensitive to significant magnitude differences (e.g., Sitting vs. Jogging). The Manhattan distance is less influenced by outliers, making it more appropriate for raw sensor data, which often contains noise. In contrast, the Chebyshev distance, defined by the maximum difference across any single feature dimension, is sensitive to extreme sensor values. The Canberra distance is particularly useful when sensor values are of different scales (e.g., accelerometer vs. gyroscope readings). Finally, the correlation distance is highly efficient when the waveform or temporal pattern of the sensor data is the main discrimination factor.

2.4. Performance Assessment

To evaluate the recognition performance of our proposed distance-based SFS-KNN model and for the sake of comparison with prior works, three principal performance metrics: accuracy, recall, and precision, were utilized. The accuracy of a classification model reflects its overall performance, whereas both recall and precision assess its performance for each particular class. Mathematically, these three evaluation metrics can be deduced from the "confusion matrix" shown in Table 5.

Table 5. A multi-class problem confusion matrix.

		Actual Classes			
		Actual Class c_1	*Actual Class* c_2	...	*Actual Class* c_n
Predicted Classes	*Predicted as* c_1	c_{11}	c_{12}	...	c_{1n}
	Predicted as c_2	c_{21}	c_{22}	...	c_{2n}
	⋮	⋮	⋮	⋮	⋮
	Predicted as c_n	c_{n1}	c_{n2}	...	c_{nn}

Fig. 2 illustrates the $n \times n$ confusion matrix, where n is the number of target classes in the dataset under consideration. For each class c_i, the matrix shows the number of positive class instances that are correctly recognized as positives (TP), the number of negative class instances that are mistakenly classified as positives (FP), and the number of positive class data samples that are incorrectly identified as negatives (FN). These terms are defined by Equations (9-11), respectively.

$$\mathrm{TP_{ci}} = \mathrm{c_{ii}} \tag{9}$$

$$\mathrm{FP}_{ci} = \sum_{k=1}^{n} c_{ik} \_ TP_{ci} \tag{10}$$

$$\mathrm{FP}_{ci} = \sum_{c_{ki}}^{n} \_ \mathrm{TP}_{ci} \tag{11}$$

The accuracy, recall, and precision of a multi-class problem are given as:

$$\mathrm{Accuracy} = \frac{\sum_{i=1}^{n} TP_{c_i}}{\sum_{k=1}^{n} \sum_{f=1}^{n} c_{kf}} \tag{12}$$

$$Average\ Recall = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_{ci}}{TP_{ci} + FN_{ci}} \tag{13}$$

$$Average\ Precision = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_{ci}}{TP_{ci} + FP_{ci}} \tag{14}$$

During feature selection and classification, standard 10-fold stratified cross-validation was used to evaluate the SFS-KNN model [57]. According to this recommended technique, the entire dataset is divided into ten subsets, each having roughly 10% of the data samples and the same class fractions as in the original set. The first nine subsets are used to train the model, whereas the remaining subset (i.e., the testing or validation set) is used to evaluate the model's performance. This process is repeated until all 10 subsets are used as validation sets.

3. Results and Discussion

Having determined the optimal feature subset for each distance function using the proposed SFS-KNN model, accuracy, precision, and recall are calculated to identify the most suitable distance for each dataset, following the aforementioned 10-fold cross-validation strategy. The classification, feature selection, and preprocessing experiments were performed using integrated tools available in Rapid Miner Studio (Educational Edition 9.10) installed on a laptop with a 2.3 GHz Intel Core i7 processor, a 256 GB SSD, and 8 GB of RAM. In this section, the results of all activity recognition experiments conducted on WISDM and UCI-HAR datasets are summarized in Tables 6 to 9. Furthermore, a detailed comparison with previous models is provided in Table 10.

Table 6. Classification results on the WISDM dataset. # F: number of selected features.

Distance	# F	Avg. Accuracy (%)	Avg. Recall (%)	Avg. Precision (%)
Manhattan	11	92.86 (±1.00)	90.03 (±1.39)	91.58 (±1.46)
Euclidean	12	93.00 (±0.92)	90.13 (±1.08)	91.59 (±1.28)
Chebyshev	10	92.12 (±1.58)	88.72 (±2.20)	90.43 (±1.88)
Canberra	7	90.38 (±0.81)	86.96 (±1.62)	88.73 (±1.21)
Correlation	16	92.21 (±0.99)	88.56 (±1.89)	90.73 (±1.65)

Table 7. Classification results on UCI-HAR dataset. # F: number of selected features.

Distance	# F	Avg. Accuracy (%)	Avg. Recall (%)	Avg. Precision (%)
Manhattan	30	99.32 (±0.37)	99.34 (±0.35)	99.35 (±0.35)
Euclidean	36	99.24 (±0.53)	99.26 (±0.49)	99.28 (±0.50)
Chebyshev	30	98.58 (±1.54)	98.58 (±1.53)	98.61 (±1.63)
Canberra	30	98.78 (±0.71)	98.80 (±0.73)	98.81 (±0.69)
Correlation	23	99.33 (±0.46)	99.35 (±0.43)	99.36 (±0.44)

Table 6 and 7 present the optimal feature count achieved by the SFS-KNN model for each distance and their relative accuracy, recall, precision ratios, along with the standard deviation values of these three performance measures for both datasets. For the WISDM dataset, it is evident that among the other four distance functions used by the proposed wrapper model, the Euclidean distance achieved the highest accuracy, recall, and precision rates of 93.39%, 90.89%, and 92.10%, respectively. Although it has produced a lower reduction ratio of 71.42%; as demonstrated by the number of selected 12 out of 42 original features; than those produced by Manhattan, Chebyshev, and Canberra, these three distances had a lower classification performance in terms of accuracy, precision, and recall. Moreover, all other four distances showed higher standard deviations values across folds than those observed for the Euclidean distance. For the UCI-HAR dataset, the wrapper model attained more than 99% accuracy using either Manhattan, Euclidean, or Correlation distances. However, the prediction results of Correlation distance were slightly higher than the other two counterparts, with 99.33% accuracy and average recall and precision of 99.35% and 99.36%, respectively. More importantly, it was able to yield a smaller feature set than the rest distance measures by reducing the original features from 561 to 23 only and achieving a greater reduction ratio of 95.9%. It is also worth mentioning that for both WISDM and UCI-HAR datasets, the proposed model had less convincing classification results based on Chebyshev and Canberra distances, indicating their unsuitability for activity recognition task. In summary, it can be noticed that:

- For the WISDM dataset, the Euclidean distance attained the highest classification performance and lower variability but with moderate reduction ratio.
- For the UCI-HAR dataset, the Correlation distance was the top performer among other distances with the best feature reduction ratio but with slightly higher variability than Manhattan and Euclidean, which came in the second place.
- For both datasets, the Chebyshev and Canberra distances exhibited less suitablility and efficiency for the HAR task using the proposed SFS-KNN model.

Table 8 and 9 display the confusion matrices achieved by the SFS-KNN approach over both human activity datasets for the two top-performing distances. They present the classification results from our proposed model for the six basic human activities in the WISDM and UCI-HAR datasets. For WISDM's confusion matrix, as indicated by the underlined values, more than 78% of classification errors (i.e., 299 out of 380 incorrectly classified instances). It arises from overlap among the three dynamic activities: *Upstairs*, *Downstairs*, and *Walking*.

Table 8. Confusion matrix of the Euclidean distance-based SFS-KNN model (WISDM dataset).

Predicted Classes	Actual Classes					
	Jogging	**Walking**	**Upstairs**	**Downstairs**	**Sitting**	**Standing**
Jogging	1592	5	15	6	0	0
Walking	7	2022	45	69	1	3
Upstairs	15	31	513	74	0	4
Downstairs	10	23	57	377	1	3
Sitting	0	0	1	1	301	3
Standing	1	0	1	1	3	233

Table 9. Confusion matrix of the Correlation distance-based SFS-KNN model (UCI-HAR dataset).

Predicted Classes	Actual Classes					
	Standing	**Sitting**	**Laying**	**Walking**	**Downstairs**	**Upstairs**
Standing	1879	28	0	0	0	0
Sitting	27	1748	0	0	0	0
Laying	0	0	1944	0	0	0
Walking	0	0	0	1720	4	3
Downstairs	0	0	0	1	1399	1
Upstairs	0	1	0	1	3	1540

This high misclassification percentage, as explained in the original work [16], is due to the analogous signal-variation patterns. Such overlap caused a significant deterioration in the classification accuracy for stair-climbing actions because both have fewer instances than the *Walking* activity. Consequently, it has reduced the overall accuracy of the proposed model, as shown in Table 6. Regarding the confusion matrix for the UCI-HAR dataset, the wrapper model shows a clear difficulty in distinguishing between the two non-dynamic actions (*Sitting* and *Standing*). The reason for this, as stated in the originally published paper [17], is the smartphone's physical position. However, the results indicate that the proposed model has successfully recognised the three dynamic actions and delivered a perfect classification for the Lying *down* activity. It should be noted that the misclassification overlaps shown in both confusion matrices are significantly less than those generated by the methods proposed in [16] and [17]. Table 10 highlights the improvement in accuracy achieved by the proposed model for each activity type, particularly for those with higher misclassification overlap in the aforementioned original studies.

Table 10. Per-activity comparison against models proposed in original works (MLP [16] and multiclass SVM [17]). Activity names are abbreviated.

Dataset	**Method**	**Overall (%)**	**% of correctly classified instances**						
			Walk	**Jog**	**Up**	**Down**	**Sit**	**Stand**	**Lie**
WISDM	MLP	91.7	91.7	98.3	61.5	44.3	95	91.9	-
	SFS-KNN (Ours)	93	97.16	97.97	81.17	71.4	98.37	94.72	-
UCI	Multiclass SVM	96	99.5	-	96	98	88.0	97	100
	SFS-KNN (Ours)	99.33	99.88	-	99.74	99.5	98.37	98.58	100

Table 10 shows that our proposed solution delivers remarkably higher discrimination results than the MLP model for the two stair-climbing activities and more robust accuracy scores for *Sitting* and *Standing* compared to multiclass SVM. It also provides much better recognition results for the set of remaining actions recorded in both datasets. This reflects the ability of our model to recognize each performed activity more efficiently. Compared with the previous optimisation methods reviewed in Section 2, our HAR model achieves the same accuracy as the best approach proposed for the WISDM dataset [23], while adopting the same feature selection method but a different classification technique. Our proposed model also achieves higher overall accuracy than all optimisation solutions for the UCI-HAR dataset, while showing performance equivalent to the method proposed in [27], with considerably better feature dimensionality reduction (23 vs 185 features).

4. Conclusion

This paper proposes an efficient wrapper-based feature selection framework to reduce feature dimensionality while achieving state-of-the-art recognition accuracy for the daily human activities. The proposed model integrates sequential forward selection technique with k-nearest neighbor classifier using five distinct distance measures (Manhattan, Euclidean, Chebyshev, Canberra, and Correlation). Two open-source and benchmark datasets, WISDM and UCI-HAR, were used to evaluate the efficiency and robusness of the proposed model. The experimental results demonstrated different feature subsets

and classification accuracies for each adopted distance. Based on Euclidean distance, the wrapper model achieved a reduction ratio of over 71% and an overall accuracy of 93% for the WISDM dataset. It also attained a significant 95.9% reduction in feature space for the UCI-HAR dataset based on Correlation distance, with a recognition rate of 99.33%. These findings confirm the superiority of our solution over previously proposed feature selection models for both datasets. For our upcoming work, we seek to use other publicly available HAR datasets, investigate larger groups of distance metrics, and combine various feature selection methods with other versions of the k-nearest neighbors algorithm to create more robust and accurate HAR solutions.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] G. Diraco, G. Rescio, A. Caroppo, A. Manni, and A. Leone, "Human Action Recognition in Smart Living Services and Applications: Context Awareness, Data Availability, Personalization, and Privacy," *Sensors*, vol. 23, no. 13, p. 6040, Jun. 2023, doi: 10.3390/s23136040.

[2] D.-Q. Vu, T. P. Thi Thu, N. Le, and J.-C. Wang, "Deep Learning for Human Action Recognition: A Comprehensive Review," *APSIPA Trans. Signal Inf. Process.*, vol. 12, no. 2, pp. 1–40, Apr. 2022, doi: 10.1561/116.00000068.

[3] H. Haresamudram, I. Essa, and T. Plötz, "Towards Learning Discrete Representations via Self-Supervision for Wearables-Based Human Activity Recognition," *Sensors*, vol. 24, no. 4, p. 1238, Feb. 2024, doi: 10.3390/s24041238.

[4] L. Boborzi, J. Decker, R. Rezaei, R. Schniepp, and M. Wuehr, "Human Activity Recognition in a Free-Living Environment Using an Ear-Worn Motion Sensor," *Sensors*, vol. 24, no. 9, p. 2665, Apr. 2024, doi: 10.3390/s24092665.

[5] V. Mahalakshmi, P. Kumar, M. Bhende, I. Keshta, S. Y. Rathod, and J. V. N. Ramesh, "IOT based wearable sensor system architecture for classifying human activity," *Meas. Sensors*, vol. 39, no. June, p. 101871, Jun. 2025, doi: 10.1016/j.measen.2025.101871.

[6] S. Javadi, D. Riboni, L. Borzì, and S. Zolfaghari, "Graph-Based Methods for Multimodal Indoor Activity Recognition: A Comprehensive Survey," *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 5, pp. 3728–3746, Oct. 2025, doi: 10.1109/TCSS.2024.3523240.

[7] K. Takenaka, K. Kondo, and T. Hasegawa, "Segment-Based Unsupervised Learning Method in Sensor-Based Human Activity Recognition," *Sensors*, vol. 23, no. 20, p. 8449, Oct. 2023, doi: 10.3390/s23208449.

[8] A. Roy, H. Dutta, A. K. Bhuyan, and S. Biswas, "On-Device Semi-Supervised Activity Detection: A New Privacy-Aware Personalized Health Monitoring Approach," *Sensors*, vol. 24, no. 14, p. 4444, Jul. 2024, doi: 10.3390/s24144444.

[9] M. Koch, T. Pfitzinger, F. Schlenke, F. Kohlmorgen, R. Groll, and H. Wöhrle, "Recognition of Human Activities Based on Ambient Audio and Vibration Data," *IEEE Access*, vol. 12, pp. 174399–174412, 2024, doi: 10.1109/ACCESS.2024.3457912.

[10] A. Grenier *et al.*, "Evaluation of Sensors Measurements in Consumer-Grade Android Wearables for Robust Positioning Applications," in *Proceedings of the International Technical Meeting of the Satellite Division of The Institute of Navigation, ION GNSS+*, Institute of Navigation, Oct. 2024, pp. 1094–1108. doi: 10.33012/2024.19675.

[11] N. A. Almujally *et al.*, "Biosensor-Driven IoT Wearables for Accurate Body Motion Tracking and Localization," *Sensors*, vol. 24, no. 10, p. 3032, May 2024, doi: 10.3390/s24103032.

[12] Y. Yin, L. Xie, Z. Jiang, F. Xiao, J. Cao, and S. Lu, "A Systematic Review of Human Activity Recognition Based on Mobile Devices: Overview, Progress and Trends," *IEEE Commun. Surv. Tutorials*, vol. 26, no. 2, pp. 890–929, 2024, doi: 10.1109/COMST.2024.3357591.

[13] V. Dentamaro, V. Gattulli, D. Impedovo, and F. Manca, "Human activity recognition with smartphone-integrated sensors: A survey," *Expert Syst. Appl.*, vol. 246, no. July, p. 123143, Jul. 2024, doi: 10.1016/j.eswa.2024.123143.

[14] X. Wang, Y. Wang, and J. Wu, "Position-Aware Indoor Human Activity Recognition Using Multisensors Embedded in Smartphones," *Sensors*, vol. 24, no. 11, p. 3367, May 2024, doi: 10.3390/s24113367.

[15] H. Jamil, M. A. Khan, and F. Jamil, "A novel hybrid strategy based on Swarm and Heterogeneous Federated Learning using model credibility awareness for activity recognition in cross-silo multistorey building," *Eng. Appl. Artif. Intell.*, vol. 138, no. December, p. 109126, Dec. 2024, doi: 10.1016/j.engappai.2024.109126.

[16] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, Mar. 2011, doi: 10.1145/1964897.1964918.

[17] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition using Smartphones," *Eur. Symp. Artif. Neural Networks*, pp. 437–442, 2013, Accessed: Aug. 05, 2026. [Online]. Available: https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf.

[18] A. Logacjov, K. Bach, A. Kongsvold, H. B. Bårdstu, and P. J. Mork, "HARTH: A Human Activity Recognition Dataset for Machine Learning," *Sensors*, vol. 21, no. 23, p. 7853, Nov. 2021, doi: 10.3390/s21237853.

[19] W. Gomaa and M. A. Khamis, "A perspective on human activity recognition from inertial motion data," *Neural Comput. Appl.*, vol. 35, no. 28, pp. 20463–20568, Oct. 2023, doi: 10.1007/s00521-023-08863-9.

[20] N. sedaghati, S. Ardebili, and A. Ghaffari, "Application of human activity/action recognition: a review," *Multimed. Tools Appl.*, vol. 84, no. 28, pp. 33475–33504, Jan. 2025, doi: 10.1007/s11042-024-20576-2.

[21] E. Zaitseva, V. Levashenko, S. Stankevich, and A. Bekenova, "Importance Analysis Method Application for Feature Selection in a Classification," in *2023 7th International Conference on System Reliability and Safety (ICSRS)*, IEEE, Nov. 2023, pp. 144–148. doi: 10.1109/ICSRS59833.2023.10381366.

[22] W. Zhang, G. Yu, and S. Deng, "Research on TCN Model Based on SSARF Feature Selection in the Field of Human Behavior Recognition," *IET Biometrics*, vol. 2024, no. 1, p. 4982277, Jan. 2024, doi: 10.1049/2024/4982277.

[23] L. Al-Frady and A. Al-Taei, "Wrapper Filter Approach for Accelerometer-Based Human Activity Recognition," *Pattern Recognit. Image Anal.*, vol. 30, no. 4, pp. 757–764, Oct. 2020, doi: 10.1134/S1054661820040033.

[24] A. M. Helmi, M. A. A. Al-qaness, A. Dahou, R. Damaševičius, T. Krilavičius, and M. A. Elaziz, "A Novel Hybrid Gradient-Based Optimizer and Grey Wolf Optimizer Feature Selection Method for Human Activity Recognition Using Smartphone Sensors," *Entropy*, vol. 23, no. 8, p. 1065, Aug. 2021, doi: 10.3390/e23081065.

[25] M. A. A. Al-qaness, A. M. Helmi, A. Dahou, and M. A. Elaziz, "The Applications of Metaheuristics for Human Activity Recognition and Fall Detection Using Wearable Sensors: A Comprehensive Analysis," *Biosensors*, vol. 12, no. 10, p. 821, Oct. 2022, doi: 10.3390/bios12100821.

[26] A. Al-Taei, M. F. Ibrahim, and N. J. Habeeb, "Optimizing the Performance of KNN Classifier for Human Activity Recognition," in *Communications in Computer and Information Science*, vol. 1440 CCIS, Springer Science and Business Media Deutschland GmbH, 2021, pp. 373–385. doi: 10.1007/978-3-030-81462-5_34.

[27] D. Thakur and S. Biswas, "Guided regularized random forest feature selection for smartphone based human activity recognition," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 7, pp. 9767–9779, Jul. 2023, doi: 10.1007/s12652-022-03862-5.

[28] P. Khatiwada, M. Subedi, A. Chatterjee, and M. Wulf Gerdes, "Automated Human Activity Recognition by Colliding Bodies Optimization-based Optimal Feature Selection with Recurrent Neural Network (RNN)," *pre prints*, vol. 1, pp. 1–17, Oct. 2020, doi: 10.20944/preprints202010.0367.v1.

[29] C. Fan and F. Gao, "Enhanced Human Activity Recognition Using Wearable Sensors via a Hybrid Feature Selection Method," *Sensors*, vol. 21, no. 19, p. 6434, Sep. 2021, doi: 10.3390/s21196434.

[30] B. A. Mohammed Hashim and R. Amutha, "Human activity recognition based on smartphone using fast feature dimensionality reduction technique," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 2, pp. 2365–2374, Feb. 2021, doi: 10.1007/s12652-020-02351-x.

[31] R. Guha, A. H. Khan, P. K. Singh, R. Sarkar, and D. Bhattacharjee, "CGA: a new feature selection model for visual human action recognition," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 5267–5286, May 2021, doi: 10.1007/s00521-020-05297-5.

[32] V. B. Semwal, P. Lalwani, M. K. Mishra, V. Bijalwan, and J. S. Chadha, "An optimized feature selection using bio-geography optimization technique for human walking activities recognition," *Computing*, vol. 103, no. 12, pp. 2893–2914, Dec. 2021, doi: 10.1007/s00607-021-01008-7.

[33] WISDM Lab, "WISDM Activity Prediction Dataset." [Online]. Available: https://www.cis.fordham.edu/wisdm/dataset.php.

[34] J. L. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra, "Human Activity Recognition Using Smartphones." [Online]. Available: https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones.

[35] B. de la Iglesia, "Evolutionary computation for feature selection in classification problems," *WIREs Data Min. Knowl. Discov.*, vol. 3, no. 6, pp. 381–407, Nov. 2013, doi: 10.1002/widm.1106.

[36] E. A. K. Zaman, A. Ahmad, and A. Mohamed, "Adaptive threshold optimisation for online feature selection using dynamic particle swarm optimisation in determining feature relevancy and redundancy," *Appl. Soft Comput.*, vol. 156, no. May, p. 111477, May 2024, doi: 10.1016/j.asoc.2024.111477.

[37] A. Hassan, J. H. Paik, S. R. Khare, and S. A. Hassan, "A wrapper feature selection approach using Markov blankets," *Pattern Recognit.*, vol. 158, no. February, p. 111069, Feb. 2025, doi: 10.1016/j.patcog.2024.111069.

[38] N. Kumar, A. Narang, and B. Lall, "KL Regularized Normalization Framework for Low Resource Tasks," in *Lecture Notes in Computer Science*, vol. 14172 LNAI, Springer Science and Business Media Deutschland GmbH, 2023, pp. 71–89. doi: 10.1007/978-3-031-43421-1_5.

[39] S. Ramadhani *et al.*, "Feature Selection Optimisation for Cancer Classification Based on Evolutionary Algorithms: An Extensive Review," *Comput. Model. Eng. Sci.*, vol. 143, no. 3, pp. 2711–2765, Jun. 2025, doi: 10.32604/cmes.2025.062709.

[40] F. Kamalov, H. Sulieman, A. Alzaatreh, M. Emarly, H. Chamlal, and M. Safaraliev, "Mathematical Methods in Feature Selection: A Review," *Mathematics*, vol. 13, no. 6, p. 996, Mar. 2025, doi: 10.3390/math13060996.

[41] Q. D. Nam Nguyen, A.-B. Liu, and C.-W. Lin, "Development of a Neurodegenerative Disease Gait Classification Algorithm Using Multiscale Sample Entropy and Machine Learning Classifiers," *Entropy*, vol. 22, no. 12, p. 1340, Nov. 2020, doi: 10.3390/e22121340.

[42] S. Shafiee, L. M. Lied, I. Burud, J. A. Dieseth, M. Alsheikh, and M. Lillemo, "Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery," *Comput. Electron. Agric.*, vol. 183, no. April, p. 106036, Apr. 2021, doi: 10.1016/j.compag.2021.106036.

[43] J. Bemister-Buffington, A. J. Wolf, S. Raschka, and L. A. Kuhn, "Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition," *Biomolecules*, vol. 10, no. 3, p. 454, Mar. 2020, doi: 10.3390/biom10030454.

[44] R. K. Sachdeva, P. Bathla, P. Rani, R. Lamba, G. S. P. Ghantasala, and I. F. Nassar, "RETRACTED ARTICLE: A novel K-nearest neighbor classifier for lung cancer disease diagnosis," *Neural Comput. Appl.*, vol. 36, no. 35, pp. 22403–22416, Dec. 2024, doi: 10.1007/s00521-024-10235-w.

[45] B. Yuan, Y. Chen, Z. Tan, W. Jinyan, H. Liu, and Y. Zhang, "Label Distribution Learning-Enhanced Dual-KNN for Text Classification," in *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2024, pp. 400–408. doi: 10.1137/1.9781611978032.47.

[46] M. Sabri *et al.*, "A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors," *J. Classif.*, vol. 41, no. 2, pp. 264–288, Jul. 2024, doi: 10.1007/s00357-024-09471-5.

[47] G. P. Kumar, K. Anbazhagan, and N. Ramasenderan, "Object classification based-on patterns using random forest classifier compared with enhanced K-nearest neighbor algorithm," in *AIP Conference Proceedings*, American Institute of Physics, Aug. 2024, p. 020200. doi: 10.1063/5.0229218.

[48] M.-N. Wang, X.-J. Xie, Z.-H. You, L. Wong, L.-P. Li, and Z.-H. Chen, "Combining K Nearest Neighbor With Nonnegative Matrix Factorization for Predicting Circrna-Disease Associations," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 20, no. 5, pp. 2610–2618, Sep. 2023, doi: 10.1109/TCBB.2022.3180903.

[49] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[50] A. Cetin and A. Buyuklu, "Revisiting distance metrics in k-nearest neighbors algorithms: Implications for sovereign country credit rating assessments," *Therm. Sci.*, vol. 28, no. 2 Part C, pp. 1905–1915, 2024, doi: 10.2298/TSCI231111008C.

[51] B.-W. Yuan *et al.*, "A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets," *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4457–4481, May 2021, doi: 10.1007/s00521-020-05256-0.

[52] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban, "Metric Learning from Imbalanced Data," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, Nov. 2019, pp. 923–930. doi: 10.1109/ICTAI.2019.00131.

[53] Z. Wang *et al.*, "Exploring global information for session-based recommendation," *Pattern Recognit.*, vol. 145, no. January, p. 109911, Jan. 2024, doi: 10.1016/j.patcog.2023.109911.

[54] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Aug. 2010, pp. 91–94. doi: 10.1109/FSKD.2010.5569740.

[55] H. A. Abu Alfeilat *et al.*, "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review," *Big Data*, vol. 7, no. 4, pp. 221–248, Dec. 2019, doi: 10.1089/big.2018.0175.

[56] R. Todeschini, D. Ballabio, and V. Consonni, "Distances and Other Dissimilarity Measures in Chemometrics," in *Encyclopedia of Analytical Chemistry*, Wiley, 2015, pp. 1–34. doi: 10.1002/9780470027318.a9438.

[57] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," vol. 1, pp. 1–17, Aug. 2020, Accessed: Aug. 05, 2026. [Online]. Available: http://arxiv.org/abs/2008.05756.