SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Discovering and Recognizing of Imbalance Human Activity in Healthcare Monitoring using Hybrid SMOTE Tomek Technique and Decision Tree Model

Raihani Mohamed[1,*], Nur Hidayah Azizan[1], Thinagaran Perumal[1], Syaifulnizam Abd Manaf[1], Erzam Marlisah[1], Medria Kusuma Dewi Hardhienata[2]

[1] Intelligent Computing RG, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2] Department of Computer Science, Faculty of Mathematics and Natural Science, IPB University, Bogor, Indonesia

ARTICLE INFO

Article history:
Received 19 June 2023
Received in revised form 17 July 2023
Accepted 23 September 2023
Available online 8 November 2023

Keywords:
Human Activity Recognition; Imbalance Data; SMOTE Tomek; Healthcare; MARDA Dataset; Decision Tree; Synthetic Minority Over-Sampling Sampling

ABSTRACT

Human activity recognition model is vital and has been use in healthcare monitoring system. Bespoke multi-modal sensors were used such as accelerometer, gyroscope, GPS, temperature, pressure mat etc. Hence, the activities involved may varied resulted on class imbalance issue therefore, the model accuracy also degraded and may not provide the desired results in all aspects. Resampling method addressed as Synthetically Minority Oversampling Technique and Tomek Link (SMOTE Tomek) is proposed to balance the target classes. Moreover, many classification algorithms such as Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT) were selected for the experiments on two datasets namely MARBLE that was publicly available and MARDA dataset. The classification accuracy achieved 98.36% with hybrid SMOTE Tomek on MARBLE dataset and 97.45% with for the MARDA dataset with total execution time 19.4ms and 42.6ms respectively. Consequently, the proposed model can be deployed in a healthcare system dashboard for effective monitoring and efficient decision making.

1. Introduction

Technology advancements have enabled the use of software systems in our daily lives. One of the in-demand systems that use the Internet of Things (IoT) and machine learning is in healthcare system. It provides a better, more intelligent, and more convenient smart environment that can assist the daily activities of humans by using intelligent sensors. Machine learning is one of the methods that is widely used in computing for data analysis, prediction, data analytics, and visualisation that generally used in healthcare or to reduce the energy consumption of a house [1,2]. This human activity recognition system (HAR) or ADLs is frequently used in the healthcare domain since we want to monitor the behaviour and health of the subjects. We can monitor the subject's behaviour at home

* *Corresponding author.*
E-mail address: raihanimohamed@upm.edu.my

https://doi.org/10.37934/araset.33.2.340350

by detecting what they are doing. Sensor-based is one of the methods used in many research papers. It seems to be the best way to monitor the activities of the residents since we can ensure that the residents' privacy is always maintained. Using cameras, for instance, will interfere with privacy issues. It is ideally suited for the elderly facing health challenges such as Alzheimer's. Monitoring activities allows other family members kept always informed about their current actions [3]. Consequently, the system ensures the prevention of any unwanted incidents and provides a constant sense of safety without them being aware [4].

The needs for accurate model for healthcare system is in demand. Hence, there are several challenges that need to be tackled to produce an effective and efficient model. As the daily activities of the residents are not distributed involving three residents reside in the same environment, it caused an imbalance of data for the classes of the target. The frequency or occurrence of different activities performed by the residents is not evenly distributed. Some activities may happen more frequently than others, creating an imbalance in the dataset. Because of the unequal distribution of activities, the dataset becomes highly imbalanced. In other words, one class (or activity) may have a significantly larger number of instances than the other classes [5]. This imbalance can lead to a bias towards the majority class during the training of a machine learning model. This bias towards the majority class can have a negative impact on the performance of the machine learning model. It may cause the model to become less effective in accurately classifying instances from the minority class since it has been exposed to a disproportionate amount of data from the majority class. As a result, the model's predictions may be inaccurate, particularly for the minority class that have less representation in the dataset [6]. To address this issue, resampling technique, Synthetically Minority Oversampling Technique and Tomek Link (SMOTE Tomek) is proposed to balance the target classes and improve the performance of the model with DT [7,8].

There are several contributions to this study. A framework consisting of proposed methods with a hybrid technique SMOTE Tomek is proposed at the pre-processing phase to cater to the unbalanced class in the real-world dataset MARDA and MARBLE [9,10]. The class prediction is proposed using the state-of-art DT to improve the model accuracy. The proposed method can automatically detect and resolve the abovementioned problems to attain overall satisfaction in the decision support system in healthcare applications. Finally, we compare the results with the several state-of-art classifiers such as DT, SVM and LR to measure its performance. The rest of this paper is organized as follows. Section 2 explores the previous related work. Meanwhile, Section 3 explains the materials and methods proposed in this study. Section 4 presents the results and discussions. Section 5 clarifies the conclusion of the overall conducted experiments.

2. Related Work

Healthcare system installed with various types of sensors and actuators such as wearable sensors (e.g., accelerometer and gyroscope) that the residents of the home will wear and ambient-based sensors (e.g., PIR, temperature sensor, and magnetic switch sensor) that will be installed on the doors, drawers, or in the rooms. All these sensors will be used to collect data on the routine activities around the home. The detection of human activity, or activities of daily living (ADLs), has been widely explored in the research area in this current era [11]. These monitoring and analysing can be beneficial to see their lifestyles suitable for elderly and dependent people so that they can continually improve their daily lives [12]. Different approaches and methods have been made in the previous works, which are the SVM model [11], the Hidden Markov model [12], and Convolutional and Recurrent Deep Learning [10]. These methods have advantages that can be beneficial in making a

classification model using machine learning for Human Activity Recognition. However, the model cannot tackle the imbalance dataset.

The issue of class imbalance is a common problem in biometrics and healthcare data. When there is an imbalance in the number of samples for different classes, it can affect the accuracy of the models. To address this issue, it is important to understand the complexity of the problem, whether it is a relatively easy or difficult case of imbalanced classes. In cases where the majority class overlaps with the minority class samples, leading to low accuracy, it is referred to as noisy data. One of the main concerns with class imbalance is the difference in sample sizes between the majority and minority classes [13]. A dataset is considered imbalanced when the class proportions are heavily skewed, with significantly more examples from some classes than others. This imbalance is commonly observed in biometrics, gene recognition, and medical datasets in real-world settings [14]. In scenarios involving binary-two class classification, the majority class refers to the negative observations that outnumber the positive (minority) observations, or vice versa [5]. These scenarios pose challenges for classification models, as they struggle to achieve high accuracy without proper treatment of the input data before the classification stage. There are several reasons for treating imbalanced class problems. Firstly, standard classifiers prioritize accuracy and may ignore the minority class. Secondly, standard classification methods assume that the data sample accurately represents the population of interest, which is not always the case in imbalanced problems. Lastly, classification methods for imbalanced issues should consider different costs associated with errors from different classes. One recent approach to addressing the class imbalance problem is through data-level techniques and approaches [15]. These methods focus on pre-processing the data to transform the imbalanced problem into balanced data by adjusting the class distribution [16].

This work focuses on addressing the problem of imbalanced classes by utilizing the Resample SMOTE Tomek method. Resampling techniques are employed to reconstruct the sample datasets, including both the training and validation sets. This approach aims to effectively utilize the collected dataset to enhance the estimation of population parameters and quantify evaluation uncertainties [17]. The Synthetic Minority Over-Sampling Technique (SMOTE) is a specific approach that is explicitly designed for "oversampling." Unlike simply duplicating existing instances, SMOTE generates new artificial examples using specific procedures. This technique has been shown to be highly useful in effectively handling imbalanced datasets [16]. This approach aims to mitigate the risk of overfitting by employing a random procedure to generate new samples. However, it is important to note that this random procedure can introduce noise or nonsensical samples. Nevertheless, SMOTE Tomek is highly regarded due to its simplicity [7]. Furthermore, this study emphasizes the importance of utilizing supervised learning methods such as DT, LR and SVM in addressing the task of predicting human activity recognition in healthcare environment.

3. Materials and Method

Figure 1 shows the proposed model for Human Activity Recognition. The dataset was collected from the environmental sensors and wearable sensors. After that, data understanding is the next stage. This stage allows us to understand the dataset and how they are being collected and written or saved in the CSV files. After understanding the dataset, it will be cleaned and go through a pre-processing process with several steps, including removing null values, data mapping, one hot encoding and feature selection. Consequently, only the training data will go through the resampling method using SMOTE Tomek. The training and testing data are then continued with feature scaling using normalization. The chosen model will be deployed in a dashboard used by the users. The details process was explained accordingly in the following sub sections.

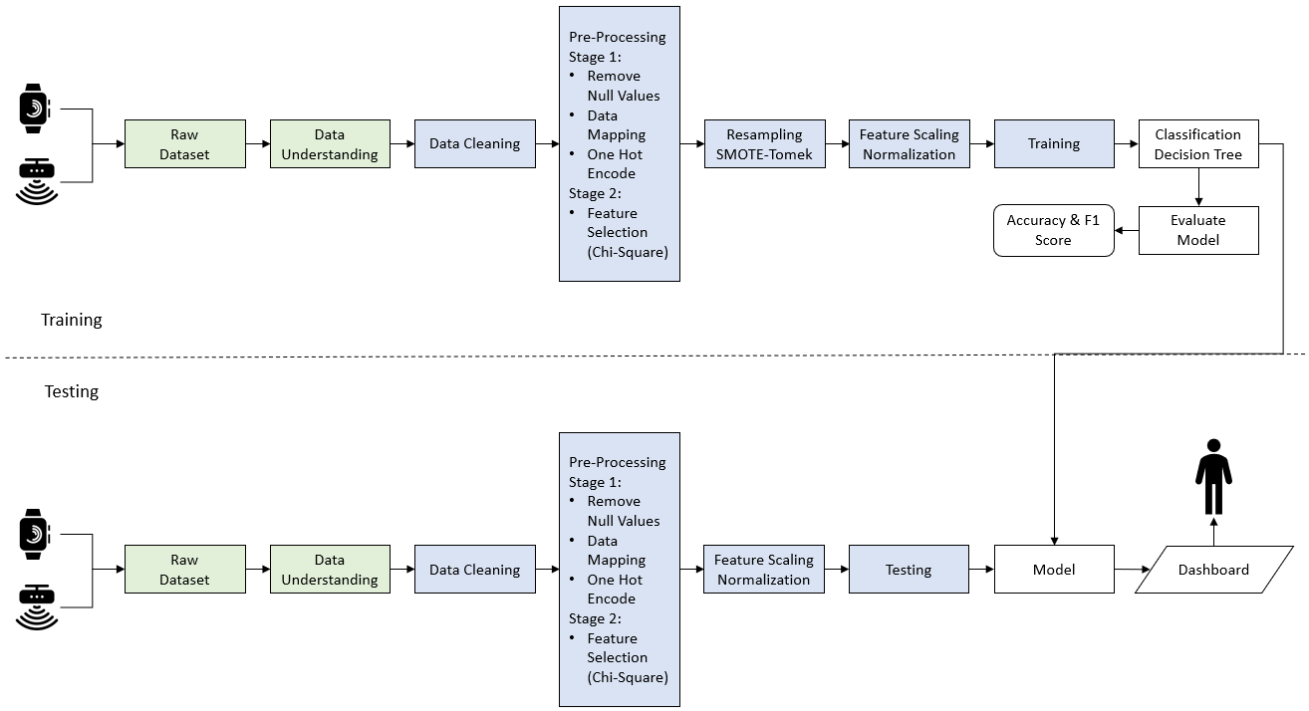


Fig. 1. Human Activity Recognition Proposed Model

3.1 Data Preparation

MARDA Dataset [9] and MARBLE dataset [10] were used to build the proposed model. Both was collected for ADLs of multi-inhabitant. This dataset is available for public use for human activity recognition activity. MARDA dataset used both wearable and environmental sensors. The wearable sensor namely accelerometer sensors were used and environmental sensors such as passive infrared, touch, analog ceramic vibration, temperature and humidity sensors were installed. Activities ADL such as cook, eat, set up table, using PC, watch TV, etc were included in 70879 rows of data and 20 attributes. MARBLE dataset is scripted; however, it has realistic scenarios suitable for modelling. Four subjects live in the same house and carry out their daily activities. A few environmental sensors were installed in the environment such as magnetic, smart plug and pressure mat. Activity executed in the smart environment including cook, eat, set up table, using PC, watch TV, wash dishes, clear table, enter home, leave home prepare meal. This dataset has 91,714 rows of data and 24 attributes.

3.2 Data Pre-Processing

3.2.1 Data cleaning and transformation

The features ON and OFF transformed into binary 1 and 0 and any activity labelled as transition is removed since it is not necessary to detect human activity. Any null values in the dataset also removed since it cannot be replaced with any values. Feature selection method also used to extract and choose only the most relevant features; hence user and timestamp also were irrelevant for the next process.

3.2.2 Resampling hybrid SMOTE Tomek method

The target of the classes was then checked to see whether they were balanced or imbalanced. After plotting a graph of the samples of the dataset for each class, as seen in Figures 2 and 3, the

datasets were seen to be unbalanced. To avoid bias in the model and improve performance, the resampling method was proposed before training the model. The resampling method used for this dataset is hybrid SMOTE Tomek. SMOTE is an oversampling method here the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together [5,6]. While Tomek Links is a down sampling method [7,18].

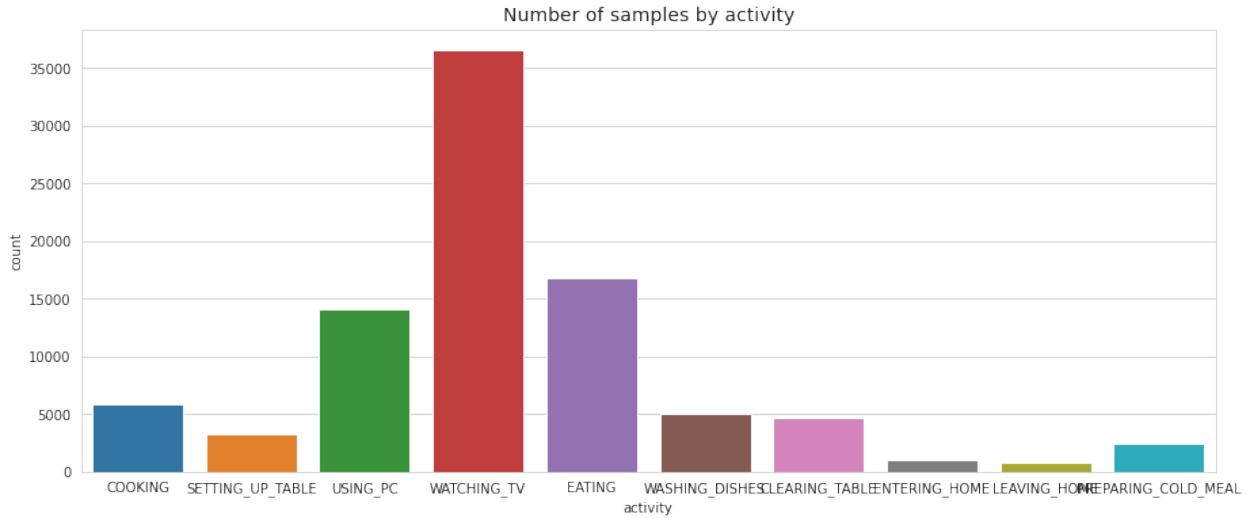


Fig. 2. Number of samples by Activity of MARBLE Dataset

Tomek developed the Tomek link, which was originally designed for two different classes (one majority and one minority), where, if the majority and minority classes are x_a and x_b, then the distance between them will be $d(x_a, x_b)$ and is known as the Tomek link, provided that no other class x_z such that $d(x_a, x_z) < d(x_a, x_b)$ or $d(x_b, x_z) < d(x_a, x_b)$ [7]. Tomek link works by eliminating the majority class instances that are closer to the minority class by applying the nearest neighbour rule to select instances and is also classified as an improved condensed nearest neighbour.

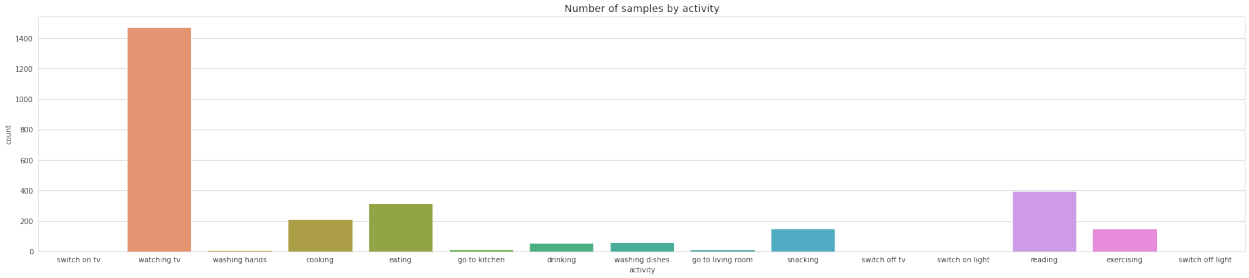
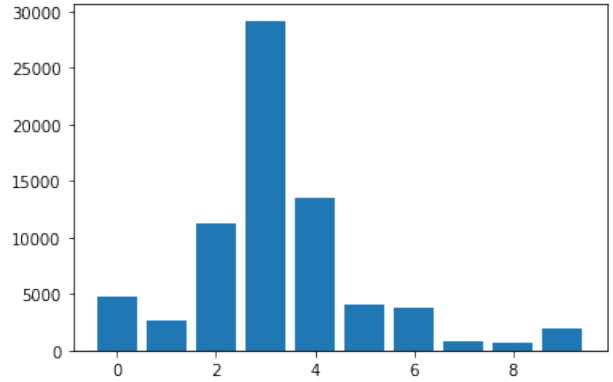


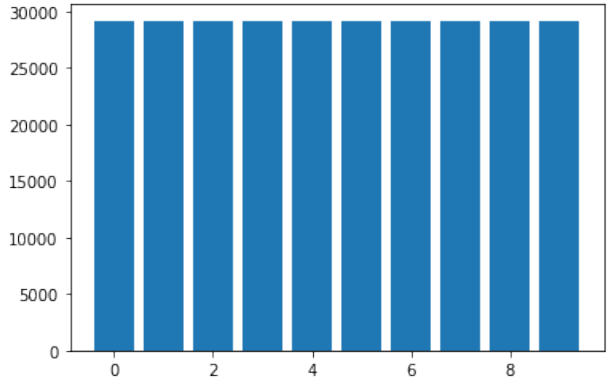
Fig. 3. Number of samples by Activity of MARDA Dataset

In brief, hybrid SMOTE Tomek technique is used in this experiment to avoid overfitting from the synthetic data made during the up-sampling method. SMOTE is a process where random data is chosen from the minority class, and then, the distance between the random data and its k-nearest neighbour is calculated. The difference will be multiplied by a random number between 0 and 1. The result will then be added to the minority class as a synthetic sample. This process is repeated until the target is balanced. Consequently, it is further processed for Tomek, a down sampling method. Tomek will choose random data from the majority class, and if its nearest neighbour is the data from the minority class, the Tomek is removed [5,7]. Before resampling, the dataset is first divided into

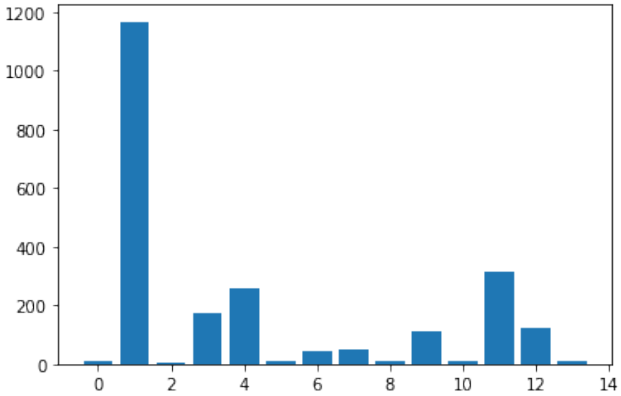
training and testing data using the train_test_split method. The training set is the only one undergoing the resampling because we want to test the model using raw data instead of resampled data. After doing the resampling method, the classes of the target can be seen to be more balanced, as we can see from Figures 4 (b) and (d) respectively.



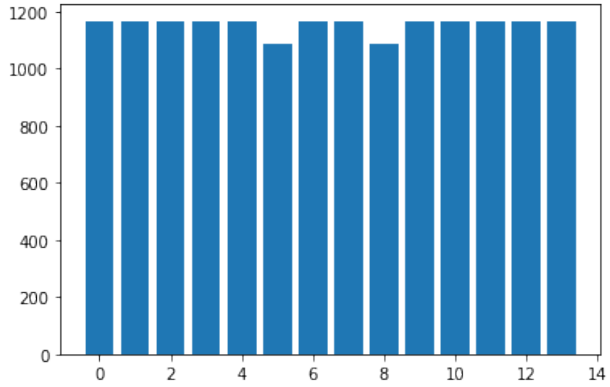
(a) Classes Distribution before Resampling process for MARBLE Dataset



(b) Classes Distribution with SMOTE Tomek method on MARBLE Dataset



(c) Classes Distribution before Resampling process for MARDA Dataset



(d) Classes Distribution with SMOTE Tomek method on MARDA Dataset

Fig. 4. Indicates class distribution of MARBLE and MARDA dataset before and after the resampling method

3.2.3 Feature scaling

Feature scaling is one of the vital processes during pre-processing stage before developing a machine-learning model. Scaling can improve a weak model and have a higher accuracy when predicting the target. The two most used feature scaling techniques are normalization and standardization. Some machine learning models are susceptible to feature scaling, while others are not. In this project, we used normalization technique where values were rescaled, so the data range between 0 and 1. This technique is also known as Min-Max scaling. Feature scaling is essential for this dataset since they have various ranges and shapes. For instance, we have labelled data from environmental sensors with values of "0" and "1" only. In contrast, data from wearable sensors like accelerometers and GPS can have values that range from negative floats to positive floats. This data has very distinct values, which may not be suitable for the machine learning model along with values “0” and “1”. Hence, normalisation has been done on the dataset to make sure the model can calculate the distances between data correctly, particularly in Logistic Regression and SVM with SGD.

3.4 Classification Methods

3.4.1 Support vector machine (SVM) and logistic regression (LR)

SVM and LR were also selected as the model classifier for these datasets. SVM is a supervised classification technique that employs hyperplanes to separate data [11]. It is a supervised machine learning algorithm that represents examples as points in space and ensures a distinct gap between categories, maximizing its width. Meanwhile LR mostly used for binary classification tasks. It works by modelling the relationship between a set of input variables (features) and a binary output variable (target) using a logistic function [19].

3.4.2 Decision tree (DT)

DT showed the best method and selected as model for this proposed method. DT is a non-parametric supervised machine learning algorithm that can be used for classification and regression problems. It has a hierarchical and tree structure which has nodes, branches, internal nodes, and leaf nodes [20]. For DT classifier parameters, the max depth is set to 12, the criterion is entropy, and the random state is 0. The same settings are set for all four experiments. For max depth, 12 is set as the value because it gives the most optimal result compared to other values. Algorithm 1 indicates the pseudocode of DT [21].

Algorithm: Decision Tree

i. Assign all training data to the root of the tree. The root of the tree will be set as the current node.
ii. For each attribute,
    - Partition all data at the node by the value of attributes.
    - Compute the information gain ratio from the partitioning.
iii. Identify feature that gives the highest gain ratio and set this feature to be the splitting criterion at the current node.
    - If the best information gain ratio is 0, set the current node as leaf and return.
iv. Partition all instances according to attribute values of the best feature.
v. Denote each partition as a child node of the current node.
vi. For each child node:
    - If the child node only has instances from one class, set is a leaf and return.
    - If not set the child node as the current node and recurse to step 2.

4. Experimental Results and Discussion

4.1 Experimental Setting

The processor used for this project is AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz with 8.00 GB Ram. This project is running on a 64-bit operating system and x64-based processor using Python, IDE Google Colab and Visual Studio Code. Two experiments were set based on objective to find the most suitable techniques and algorithms for the classification model. For the first experiment, the data is split into a 70:30 ratio without resampling. The class target is not adjusted, so the target classes are imbalanced. 70% of the overall data is used as the training input of the model, while another 30% is used as the testing set for the input. While, second experiment the data is split into a 70:30 ratio as well. 70% of the data is put aside as a training set of input, while 30% is

used as a testing set of data input. However, the data has been resampled before it is used as the training data. The data has been resampled using SMOTE Tomek technique which is an oversampling and under sampling technique. Different parameters have been set for different algorithms. For the DT classifier, the criterion is entropy, the random state is 0, and the max depth is 12. Other than that, the parameters for SVM with SGD are hinge for loss, 12 as penalty and max iteration are 5. Finally, logistic regression uses the default parameters. These are the parameters and settings for the MARBLE dataset. For the MARDA dataset, the parameters of the DT are entropy as the criterion, 0 as the random state value, and the max depth is 12. SVM with SGD, the loss is set as the hinge, the penalty is 12, and the max iteration is 5. Lastly, the parameter for logistic regression is the same as the default. All of these parameters have been tuned as the most optimal values and SMOTE Tomek is used as the default parameter.

4.2 Results Based on Experimental Settings

The model's results and output will be evaluated and discussed in this section based the accuracy and F1 Score. Four experiments and its results were showed in Table 1. Result of experiment 1 in Table 1 shows the result with a 70:30 ratio without resampling, while the result of experiment 2, using a 70:30 ratio with SMOTE Tomek. According to these two experiments, experiment 2 has a better overall performance since the accuracy and F1 Score is higher than experiment 1.

Table 1
Model results based on 4 types of experimental settings

Setting	Result (%)	Decision Tree	Logistic Regression	SVM with SGD
Experimental 1 without SMOTE Tomek	Accuracy	97.28	59.45	49.30
	F1 Score	94.00	31.00	27.00
Experimental 2 with SMOTE Tomek	Accuracy	98.36	90.79	97.75
	F1 Score	97.00	84.00	96.00

The overall performance result for the four experiments indicates that model with resampled data is much better than the dataset without resampling. Although DT have high accuracy without resampling, this is due to the bias of the classification model when making a prediction. These two algorithms tend to assign the data to the majority classes rather than the other minority classes. This will cause the classification model to be biased towards majority classes and produce an inaccurate prediction value. Thus, when training the model using the imbalanced dataset, the model will not be able to predict the correct value when using future unseen data. We can see that the accuracy and F1 score of the SVM with SGD and LR before resampling were very low. However, after applying hybrid SMOTE Tomek's resampling technique to the training data, we can see that all algorithms produce higher and better accuracies.

The best model was used and test on MARBLE datasets and provide its result as in Table 3. In term of execution time, Table 2 presents the execution time for the MARBLE and MARDA datasets.

Table 2
Execution Time Results

Dataset	Execution Time (ms)		
	Training	Testing	Total
MARBLE	6.5	12.9	19.4
MARDA	41.3	1.3	42.6

4.3 Comparison with Benchmarked Model

This section discusses the comparison between previous works and the proposed method. The significant difference of the proposed method with the earlier works as in Table 3 is the methods that are proposed different based on the objective of each study on a smart environment using the same types of sensors that was accelerometer and gyroscope [11,12]. We also compared model proposed by the author [10] using same MARBLE dataset to achieve an accuracy model based on the method proposed respectively. The highest accuracy from previous models is from Arrotta *et al.,* which is 90.32%. However, the proposed model using hybrid SMOTE Tomek technique and DT showed highest accuracy model using MARDA with 98.36% accuracy and 97.45% accuracy on MARBLE dataset.

Table 3
Method Comparison with previous works

Method	Dataset	Accuracy (%)
SVM [11]	CASAS	84.00
Hidden Markov Model [12]	ARAS	76.20
CNN-LSTM [10]	MARBLE	90.32
MARBLE Dataset using DT with SMOTE Tomek	MARBLE	97.45
MARDA Dataset using DT with SMOTE Tomek	MARDA	98.36

5. Conclusions

Human activity recognition is significant and valuable in this technological era. It is advantageous and beneficial for elderlies or any individuals who may require constant health monitoring. For instance, people with Alzheimer's or seizures and living alone may need this system installed in their homes. Other than that, pre-processing is a crucial stage in developing a machine learning model to make sure the model can understand the data and learn well. Balanced data is fundamental as it will determine how the model will learn and produce a reliable and efficient model. Imbalanced data can cause the model to falsely learn the pattern of the data and make a false prediction which can pose significant risk. Thus, preparing and giving relevant data for the model to learn beforehand is vital. In this study, a machine learning model has been made to find the best method with the highest accuracy. The hybrid SMOTE Tomek method is proposed to cater to the problem mentioned is well defined for the domain in ensuring the decision support system is deliverables and beyond the expectation of its end-user. A machine learning model using DT has been used in this study and this model can be used using live data in an intelligent system.

6. Data Availability

The datasets used in the article are publicly available standard benchmark datasets referred to in Refs. [9,10].

Acknowledgement
This work has been sponsored by IPM Putra Research Grant Scheme (Vote Number: 9699600) Universiti Putra Malaysia funded by Malaysian Ministry of Higher Education.

References

[1] Mohamed, Raihani, Thinagaran Perumal, MdNasir Sulaiman, and Norwati Mustapha. "Multi-resident activity recognition using label combination approach in smart home environment." In *2017 IEEE International Symposium on Consumer Electronics (ISCE)*, pp. 69-71. IEEE, 2017. https://doi.org/10.1109/ISCE.2017.8355551

[2] Jahromi, Ali Jamali, Mohammad Mohammadi, Shahabodin Afrasiabi, Mousa Afrasiabi, and Jamshid Aghaei. "Probability density functio n forecasting of residential electric vehicles charging profile." *Applied Energy* 323 (2022): 119616. https://doi.org/10.1016/j.apenergy.2022.119616

[3] Alberdi, Ane, Alyssa Weakley, Maureen Schmitter-Edgecombe, Diane J. Cook, Asier Aztiria, Adrian Basarab, and Maitane Barrenechea. "Smart home-based prediction of multidomain symptoms related to Alzheimer's disease." *IEEE journal of biomedical and health informatics* 22, no. 6 (2018): 1720-1731. https://doi.org/10.1109/JBHI.2018.2798062

[4] Ivașcu, Todor, and Viorel Negru. "Activity-Aware Vital Sign Monitoring Based on a Multi-Agent Architecture." *Sensors* 21, no. 12 (2021): 4181. https://doi.org/10.3390/s21124181

[5] Mohamed, Raihani, Abdul Raziff, Aabdul Rafiez and Mohd Nasir, Sabri. "A RESAMPLE-SMOTE BALANCE WITH RANDOM FOREST FOR IMPROVING SEMINAL QUALITY PREDICTION IN HEALTHCARE INFORMATICS," *ARPN Journal of Engineering and Applied Sciences*, vol. 16, no. 21, pp. 2264–2274, (2021). http://www.arpnjournals.org/jeas/research_papers/rp_2021/jeas_1121_8740.pdf

[6] Kumar, Pradeep, Roheet Bhatnagar, Kuntal Gaur, and Anurag Bhatnagar. "Classification of imbalanced data: review of methods and applications." In *IOP conference series: materials science and engineering*, vol. 1099, no. 1, p. 012077. IOP Publishing, 2021. https://doi.org/10.1088/1757-899X/1099/1/012077

[7] Swana, Elsie Fezeka, Wesley Doorsamy, and Pitshou Bokoro. "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset." *Sensors* 22, no. 9 (2022): 3246. https://doi.org/10.3390/s22093246

[8] San, Gan Shu, Siana Halim, Debora Anne Yang Aysia, and Jessie Lestari. "Predicting Willingness to Donate Smartphones as a Reuse Option Using Decision Tree Analysis." PhD diss., Petra Christian University, 2023.

[9] Mohamed, Raihani, Perumal, Thinagaran, S. N. Mohd Rum, and U. I. Mohd Izni. "MARDA: Design and Model Three Residents' Activity Recognition for Healthcare Monitoring with Machine Learning," in *Book Chapter for International Symposium in Computer Science for Smart Agriculture, Education and Medical: A Perspective from Indonesia and Malaysia*, Book Chapt. IPB Press, (2023), pp. 1–8.

[10] Arrotta, Luca, Claudio Bettini, and Gabriele Civitarese. "The marble dataset: Multi-inhabitant activities of daily living combining wearable and environmental sensors data." In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 451-468. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-94822-1_25

[11] Cook, Diane J., Aaron S. Crandall, Brian L. Thomas, and Narayanan C. Krishnan. "CASAS: A smart home in a box." *Computer* 46, no. 7 (2012): 62-69. https://doi.org/10.1109/MC.2012.328

[12] Alemdar, Hande, Halil Ertan, Ozlem Durmaz Incel, and Cem Ersoy. "ARAS human activity datasets in multiple homes with multiple residents." In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 232-235. IEEE, 2013. https://doi.org/10.4108/icst.pervasivehealth.2013.252120

[13] Wang, Hong, Qingsong Xu, and Lifeng Zhou. "Seminal quality prediction using clustering-based decision forests." *Algorithms* 7, no. 3 (2014): 405-417. https://doi.org/10.3390/a7030405

[14] Liang, Xiaomin, Daifeng Li, Min Song, Andrew Madden, Ying Ding, and Yi Bu. "Predicting biomedical relationships using the knowledge and graph embedding cascade model." *PLoS One* 14, no. 6 (2019): e0218264. https://doi.org/10.1371/journal.pone.0218264

[15] Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, no. 4 (2011): 463-484. https://doi.org/10.1109/TSMCC.2011.2161285

[16] Díez-Pastor, José F., Juan J. Rodriguez, Cesar Garcia-Osorio, and Ludmila I. Kuncheva. "Random balance: ensembles of variable priors classifiers for imbalanced data." *Knowledge-Based Systems* 85 (2015): 96-111. https://doi.org/10.1016/j.knosys.2015.04.022

[17] Charte, Francisco, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation." *Knowledge-Based Systems* 89 (2015): 385-397. https://doi.org/10.1016/j.knosys.2015.07.019

[18] Tomek, Ivan. "Two modifications of CNN." (1976).

[19] Abdellaoui, Mehrez, and Ali Douik. "Human Action Recognition in Video Sequences Using Deep Belief Networks." *Traitement du Signal* 37, no. 1 (2020). https://doi.org/10.18280/ts.370105

[20] Eloudi, Hasna, Mohammed Hssaisoune, Hanane Reddad, Mustapha Namous, Maryem Ismaili, Samira Krimissa, Mustapha Ouayah, and Lhoussaine Bouchaou. "Robustness of Optimized Decision Tree-Based Machine Learning Models to Map Gully Erosion Vulnerability." *Soil Systems* 7, no. 2 (2023): 50. https://doi.org/10.3390/soilsystems7020050

[21] Rahman, Muhammad Muhitur, Md Shafiullah, Md Shafiul Alam, Mohammad Shahedur Rahman, Mohammed Ahmed Alsanad, Mohammed Monirul Islam, Md Kamrul Islam, and Syed Masiur Rahman. "Decision Tree-Based Ensemble Model for Predicting National Greenhouse Gas Emissions in Saudi Arabia." *Applied Sciences* 13, no. 6 (2023): 3832. https://doi.org/10.3390/app13063832