European Journal of Molecular & Clinical Medicine

ISSN 2515-8260 Volume 08, Issue 04, 2021

A random forest-based class imbalance analysis in Nurse Care Activity

[1]C. Vasantha Kumari,

Assistant Professor Department of Medical Surgical Nursing Sri Venkateswara College of Nursing, Chittoor – 517127, AP

[2]P. Mohana Priya,

Associate Professor Department of Child Health Nursing, Sri Venkateswara College of Nursing, Chittoor – 517127, AP

[3]Prof. Edna Sweenie J,

Deputy Director & Professor, Department of Child Health Nursing, Sri Venkateswara College of Nursing, Chittoor – 517127, AP

[4]T. Gayathri,

Professor Department of Medical Surgical Nursing, Sri Venkateswara College of Nursing, Chittoor – 517127, AP

[5]S. Sujitha,

Associate Professor Department of Child Health Nursing, Sri Venkateswara College of Nursing, Chittoor – 517127, AP

Abstract- Because nurse care activity identification has a high class imbalance issue and intra-class variability depending on both the subject and the receiver, it is a novel and demanding study topic in human activity recognition (HAR). To address the issue of class imbalance in the Heiseikai data, nurse care activity dataset, we used the Random Forest-based resampling approach. A Gini impurity-based feature selection, model training, and validation using Stratified KFold cross-validation are all part of this technique. Random Forest classification yielded 65.9 percent average cross-validation accuracy in categorising 12 tasks performed by nurses in both laboratory and real-world contexts.. This algorithmic pipeline was created by the "Britter Baire" team for the "2nd Nurse Care Activity Recognition Challenge Using Lab and Field Data."

Keywords: Activity recognition; Nurse care; Accelerometer; Feature selection; Stratified KFold cross-validation; Random Forest.

I. INTRODUCTION

Human activity recognition has emerged as a prominent issue in active research during the last several years (HAR).

Machine Learning, Machine Perception, Artificial Intelligence, Ubiquitous Computing and Human-Computer Interaction are just few of the fields that fall under this umbrella. The goal of activity recognition is to identify a person's activities based on observations of the person and the environment around them. With HAR, people's daily routines may be observed via the analysis of data acquired from a variety of sensors on them and their immediate surroundings.

Remote monitoring of patients' activities or the activities of old individuals at home is the primary focus of this research in health care applications. However, certain nurses' actions in the hospitals are overlooked, which might have several benefits, such as automated record production, monitoring compliance with care routines for a specific patient, and identifying risk behaviours that need special attention, among others.. It's a difficult area of study since, in contrast to other forms of activity recognition in which users conduct an action on their own, nurses perform the majority of actions on patients. Intra-class variability arises as a result of this, which is influenced by both the subject and the patient getting treatment. The goal of "The 2nd Nurse Care Activity Recognition Challenge Using Lab and Field Data" is to investigate the feasibility and limitations of adopting activity recognition using movement and location in this field.

Nurses are tasked with completing 12 tasks in the lab and in the real world as part of this assignment. There is a significant probability of missing labels during studies in the actual world since nurses are always at work. Data from both training and testing environments are included in training data, however testing data is derived only from data obtained in the actual world.

The most difficult part of this project is using laboratory data to develop real-world models and bridging the model gap. That's why, in Section 2, relevant activity recognition works are recognised, and then important works and research needs are given. Heiseikai data, nurse care activity data set are briefly discussed in Section 3. Section 4 explains our suggested technique in great detail. Problems and difficulties associated with this data collection are discussed in this section. In addition, the measures performed as a result of these circumstances are briefly described. Section 5 presents the outcomes of several algorithms. In addition, we do a thorough analysis of the data using the model we selected. Section 6 concludes the paper with suggestions for further research.

II. RELATEDWORKS

In the field of activity recognition, there have been several studies, particularly in the area of mobile activity recognition. There are, however, relatively few studies on the identification of nursing activities. This year, a challenge to recognise nursing activities was organised and several teams engaged in it, with encouraging outcomes. Different characteristics and basic classifiers, such KNN, were suggested.

Motion capture and meditag sensors were used. Skeletons may be represented using a spatiotemporal graph, which was introduced. Aside from voice recognition and natural language

processing, Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) may also be employed in activity detection since they are sequential data. Mobile activity recognition might benefit from an approach known as "bag of features". Creates histograms from feature-extracted data by using that data as input. This strategy, however, necessitates the division of data. Data gathered in a controlled laboratory environment has poor performance in real-world situations. There are a wide range of activity durations and class imbalances in real-world situations, both of which may impede machine learning algorithms. For machine learning algorithms, over fitting is a typical concern. It's a good idea to use Random Forest to prevent over fitting your model. An ensemble of decision trees is known as a Random Forest. Bootstrapping is the process of creating multiple decision trees from a small portion of a dataset. After then, make judgments based on the remaining information and follow the results of a vote by the people. Bagging is the name given to this procedure. It has decent generalization ability and is less prone to over fitting since it doesn't operate with the whole data set at once and takes decisions from multiple randomly generated forests. Random Forest has been used in a number of studies on human activity identification. Using a Random Forest method, the authors of were able to correctly categorise 93.44 percent of human activities. Using a linear forward feature selection technique and a Random Forest-based methodology.

Table 1: Nurse Care Activity Data Set

Principal Category	Activity Name (Principal)	Label in Dataset	Activity Name (Sub)
A	Help in Mobility	1	Guide (from the front)
		2	Partial Assistance
		3	Walker
		4	Wheelchair
B	Assistance in Transfer	5	All Assistance
		6	Partial Assistance (from the front)
		7	Partial Assistance (from the side)
		8	Partial Assistance (from the back)
C	Position Change	9	To Supine Position /To Right Lying Position
		10	To Left Lying Position
		11	Lower Body Lifting
		12	Horizontal Movement

It is difficult for machine learning algorithms to classify data that is unbalanced. This issue has been addressed in the literature. Class imbalance is a prevalent concern in the medical field. Many diseases are only seen briefly, and as a result there may be few samples available. Other successful strategies for dealing with uneven data include up sampling the minority class and down sampling the majority class, using a high cost on the majority class and a low cost on the minority class. [We opted to use Random Forest for this dataset since it is so lopsided based on

all of previous studies. In contrast to these approaches, ours uses both up sampling and down sampling, as well as oversampling prior to filtering, this is a major distinction between the two.

III. DATA SET

The Kyushu Institute of Technology's Smart Life Care Unit in Japan gathered the test results. Data was gathered from a Japanese care facility. In the lab, 2 people participated who are professional nurses. In the actual world, 47 people took part in the experiment, however only the training and testing data of six nurses and three test nurses are included in this challenge. The data set consists of care actions that nurses undertake at the Care facility. Mobility support, transfer assistance, and position change are the three main categories of activities. Table 1 shows the many subcategories of these activities.

To obtain the data, the accelerometer sensor in the smartphone and motion capture sensor was utilised however only the accelerometer data is accessible for this challenge. The wristband that held the smartphone in place was worn on the right arm. The data is sampled at a rate of 60 hertz (Hz). No pre-processing procedure is performed to this data. The field data collection, in particular, has a large number of unlabeled observations. Also, the activity labels in this dataset are heavily skewed. Each segment of the test data set must be predicted by an activity id model that incorporates both the lab and field data sets.

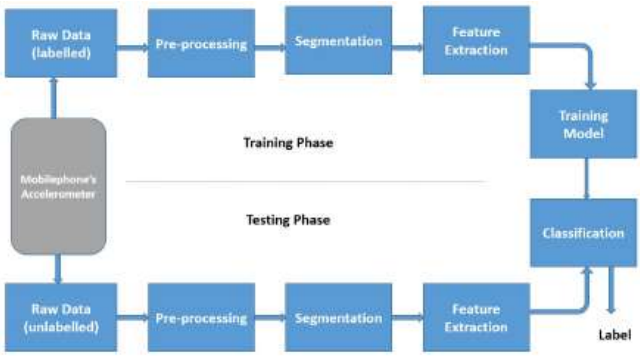

Figure 1: Basic Structure of Nurse Care Activity Recognition System.

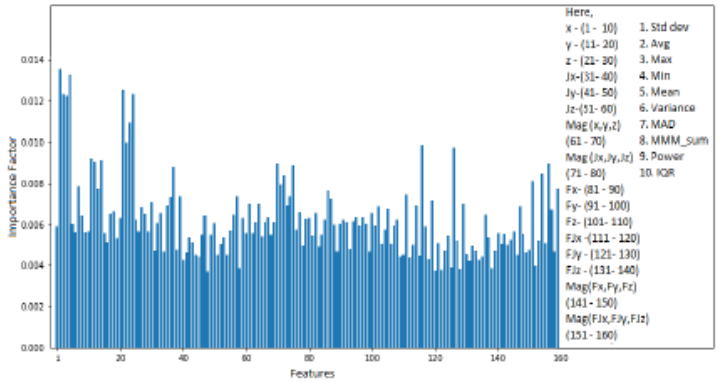

Figure 2: Feature Importance Plot.

IV. METHODOLOGY

Raw sensor data may be analysed in a number of methods that are well-established in the scientific community. According to Fig. 1, this process begins with pre-processing raw sensor data, which includes segmentation, feature extraction, and classification.

Body acceleration, gravitational acceleration, and noise are all components of a single accelerometer signal. Noise filtering and the separation of body and gravity acceleration are necessary before the segmentation approach can be used. Given that this data set consists only of body acceleration, the pre-processing step employs a low pass Butterworth filter with a cutoff frequency of 20 Hz in order to remove any noise. Different segmentation approaches may be used to extract meaningful information from a continuous stream of data in order to enhance the relevant features of the signal. The sensor signal must first be broken down into smaller time intervals known as windows. A sliding window approach, with windows of constant length and zero overlap, was utilised to split the signal in this case. There is little to no pre-processing required for this windowing technology, which makes it ideal for real-time applications and easy to implement.

A triaxial accelerometer was employed in this experiment. A three-dimensional axis model is used to visualise the data. We used the triaxial data to derive jerk, magnitude, and frequency domain signals to better identify the various nursing actions. So, the data set now has 16 input columns: x, y, z, Jx, Jy, Jz, Mag(x, y, z), Mag(Jx, Jy, Jz), Fx, Fy, Fz, FJx, FJy, FJz, Mag(Fx, Fy, Fz), Mag(Jx, Jy, Jz) (FJx, FJy, FJz).

As a result of these 16 columns, we calculated the mean, median absolute deviation (MAD), weighted average (WA), and standard deviation of 10 characteristics, including energy and IQR. As a result, there are 160 vectors in the feature space.

An essential part of activity identification and analysis is the selection of features. Every characteristic is not equally important, and some variables may be completely unnecessary. Using feature selection to prevent overfitting, increase accuracy, and decrease training time is a three-pronged approach to reducing the amount of data needed to train an algorithm.

On the basis of Gini impurity, we narrowed down the key characteristics. Fig. 2 illustrates the relative relevance of all the characteristics by way of a bar chart. Here, the percentage significance of the most essential qualities is shown against each feature's relative relevance. In order to rank and compare characteristics, this importance is determined for each individual attribute in the dataset. Attribute split points are weighted according to the number of observations they are responsible for in order to determine the relevance of a particular tree. The Gini index is used to determine the points at which to divide the sample.

All of the decision trees in the model are then averaged to get a total of the feature significance factors. Finally, 72 features were chosen among 160 characteristics based on their relevance. These results show that the top 10 attributes are as follows: average, mean, standard deviation, minimum, maximum, minimum, maximum, avg(z), variance, and variance (FJy).

We see data as a precious resource, and we want to make full use of it. Only the piece of data designated for training may be used for model training if we use the train-test split. As the quantity of training data grows, so does the quality of the models. Cross-validation is a possible solution to this problem. Data sets are partitioned into N subsets using cross-validation. During training, the N-1 split is employed, with the remainder of the split being used for testing. Using a different split each time, the model goes over the data set N times.

For both training and testing purposes, we utilise all of the data points available.

A model's performance may also be improved by using cross-validation to test it on additional, previously untested data points. KFold and Stratified KFold are two of the most often used algorithms for dividing data in cross-validation. While cross-validation is still used, the class distribution within the data set is retained in both training and testing splits when using the Stratified KFold approach.

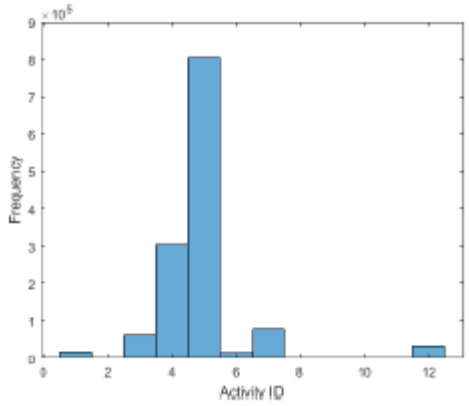

Figure 3: Field Data Histogram Plot

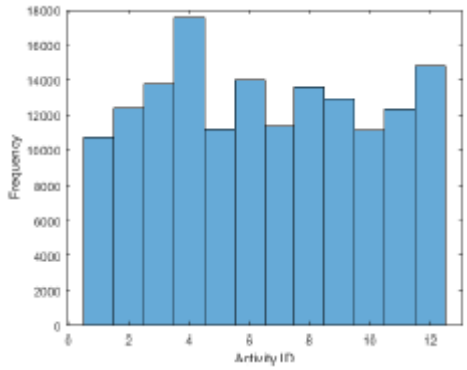

Figure 4: Lab Data Histogram Plot

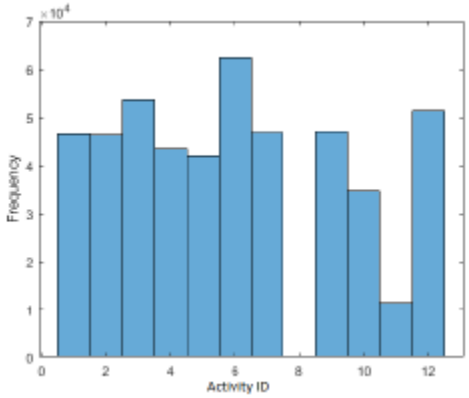

Figure 5: Field Data Histogram Plot (After Resampling)

When there is a large amount of data, Stratified KFold performs better than KFold. This can be seen in the histogram plots of field and lab data in Figure 3 and Figure 4. Resampling is used to keep things under check. If we compare Figs. 3 and 4 to Figs. 5, because there was no data for activity 8 and just a tiny quantity of data for activity 11, this occurrence occurred. As a consequence, oversampling in the field data had little to no reference point, resulting in an obvious class imbalance in the combined data. So, Stratified KFold is used in our study since it attempts to preserve the proportion of data that is present in the train-test split. This will prevent any kind of partiality. In this case, we employed cross-validation with 10 folds to improve speed.

MATLAB Statistics and Machine Learning Toolbox were used to assess the classifier's performance. The Random Forest classifier was estimated to have the best performance. Adding insult to injury, the data set is very unbalanced, which makes it extremely susceptible to overfitting. Overfitting may be readily reduced using the Random Forest classifier. Following are the random search hyperparameters for the Random Forest classifier: The number of trees is 1800, the minimum number of splits is 2, the minimum number of leaves is 1, the maximum number of features is the square root of the total number of features, and bootstrap is True.

Table 2: Accuracy Comparison of all the Tested Models

Window size	Classifier	Accuracy (%)
128	Random Forest	65.9
	Cubic SVM	57.8
	Quadratic SVM	52.3
	Fine Tree	41.2
	Medium Tree	33.6
	Linear Discriminant	33.5
	Fine KNN	55.7
	Boosted Trees	36.6
256	Random Forest	63.2
	Cubic SVM	56.8
	Quadratic SVM	52.4
	Fine Tree	46.7
	Medium Tree	47.6
512	Random Forest	63.9
	Cubic SVM	57.7
	Fine Tree	48.9
	Boosted Trees	47.7
	Cosine KNN	48.8

V. RESULT AND ANALYSIS

We tried a variety of window widths, feature sets, and algorithms to find the most effective solution in this project. All of the evaluation's experimental findings will be summarised in this section. Initial tests were conducted using MATLAB Statistics and Machine Learning Toolbox to predict the best potential combination. Table 2 shows the best potential results from all of the data. In every section, we can observe that the Random Forest Classifier has provided the greatest accuracy if we look into the data attentively.

There is a considerable risk of bias and overfitting since our data set is so uneven. With the correct hyperparameters and Random Forest classifier with Stratified KFold cross-validation, we can easily minimise this chance.

If you're dealing with an unbalanced class situation, you can't utilise accuracy as a performance metric.

In order to obtain a better sense of how our suggested strategy performs, we have employed different performance indicators, such as precision, recall, and F1 score. We have prioritised precision, recall, and F1 score above accuracy because to the data set's extreme class imbalance. The resampled data had a maximum accuracy of 65.9%. Table 3 shows the findings for both resampling and non-resampling. Low false-positive rates are associated with high accuracy. With resampling, the average precision is 0.67, whereas without resampling, it is 0.69. With resampling, we've seen an average recall of 0.66, which is over the 0.5 threshold for this model. However, without resampling, the average recall score is only 0.41. For classes with a more unequal distribution of students, the F1 score is typically more valuable than the accuracy score. Resampled and unresampled data accuracy is fairly similar in our situation, but their average F1

scores are considerably different: 0.66 for resampled data and 0.50 for unresampled data. Fig. 9 depicts the proposed classifier's confusion matrix, which demonstrates that Wheelchair and All A are the most puzzling activities.

Right Lying (LL) and Supine position (SP/RL) are the second most perplexing activities (LL).

VI. CONCLUSION AND FUTUREWORK

Class imbalance, missing labels, missing points, and incorrect timestamps were among the many issues we encountered when working with this data.

Labeling errors made by humans are mostly to blame for these issues.

With the Random Forest Classifier, we have performed Stratified KFold cross-validation in an attempt to minimise overfitting due to class bias. The Random Forest classifier provides the greatest result, with an accuracy rate of 65.9%, when accuracy is compared using several methods. This model has a lot of room for growth, as seen by an examination of the data for each performance parameter. Even after resampling, there is still a substantial class imbalance in the field data, which is why the expected result was not achieved. Data augmentation and resampling are two options to consider. During the course of our study, this approach yielded an accuracy of 70.85 percent. Resampling and data augmentation are often employed to solve this sort of issue, however in our instance they didn't work well together. In light of the rarity of this data collection, further experimentation and analysis are necessary before the concepts of resampling and data augmentation can be used here. Due to a lack of time, more testing has not yet been performed and will be included in our next project. Semi-supervised learning may also be used to produce pseudo labels and data points if there are too many missing data and labels. As a result, there will be less of a wealth gap. The lack of computing resources prevented us from implementing this technique, however others who will deal with this data set in the future may use this method. The testing dataset's recognition result will be included in the challenge summary article.

REFERENCES

[1] Rahman, A., Hassan, I., & Ahad, M. A. R. (2021, September). Nurse Care Activity Recognition: A Cost-Sensitive Ensemble Approach to Handle Imbalanced Class Problem in the Wild. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 440-445).

[2] Alia, S. S., Adachi, K., Hossain, T., Le, N. T., Kaneko, H., Lago, P., ... & Inoue, S. (2021, September). Summary of the Third Nurse Care Activity Recognition Challenge-Can We Do from the Field Data?. In Adjunct Proceedings of the 2021 ACM International

Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 428-433).

[3] Matsuyama, H., Yoshida, T., Hayashida, N., Fukushima, Y., Yonezawa, T., & Kawaguchi, N. (2020, September). Nurse care activity recognition challenge: a comparative verification of multiple preprocessing approaches. In Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (pp. 414-418).

[4] Basak, P., Tasin, S. M., Tapotee, M. I., Sheikh, M. M., Sakib, A. N., Baray, S. B., & Ahad, M. A. R. (2020, September). Complex nurse care activity recognition using statistical features. In Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (pp. 384-389).

[5] Kowshik, M. A., Pritom, Y. A., Rahman, M. S., Akbar, A., & Ahad, M. A. R. (2021, September). Nurse Care Activity Recognition from Accelerometer Sensor Data Using Fourier-and Wavelet-based Features. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 434-439).

[6] Kadir, M. E., Akash, P. S., Sharmin, S., Ali, A. A., & Shoyaib, M. (2019, September). Can a simple approach identify complex nurse care activity?. In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (pp. 736-740).

[7] Lago, P., Alia, S. S., Takeda, S., Mairittha, T., Mairittha, N., Faiz, F., ... & Inoue, S. (2019, September). Nurse care activity recognition challenge: summary and results. In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (pp. 746-751).

[8] Anand, S. J., Magesh, S., & Arockiamary, I. (2021). Contemporary Human Activity Recognition Based Predictions by Sensors Using Random Forest Classifier. Journal of Computational and Theoretical Nanoscience, 18(4), 1243-1250.

[9] Anand, S. J., Magesh, S., & Arockiamary, I. (2021). Contemporary Human Activity Recognition Based Predictions by Sensors Using Random Forest Classifier. Journal of Computational and Theoretical Nanoscience, 18(4), 1243-1250.

[10] Wang, J., Chen, Y., Gu, Y., Xiao, Y., & Pan, H. (2018, July). SensoryGANs: An effective generative adversarial framework for sensor-based human activity recognition. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.