A Consensus Novelty Detection Ensemble Approach for Anomaly Detection in Activities of Daily Living

Salisu Wada Yahaya, Ahmad Lotfi and Mufti Mahmud

School of Science and Technology, Nottingham Trent University
Nottingham, NG11 8NS, United Kingdom
Email: {salisu.yahaya2015, ahmad.lotfi, mufti.mahmud}@ntu.ac.uk

Abstract

A new approach to creating an ensemble of novelty detection algorithms is proposed in this paper. The novelty detection process identifies new or unknown data by detecting if a test data differs significantly from the data available during training. It is applicable for anomaly detection in a situation where there is sufficiently large training data representing the normal class and little or no training data for the anomalous (abnormal) class. Abnormality in Activities of Daily Living (ADL) is identified as any significant deviation from an individual's usual behavioural routine. Novelty detection is relevant to ADL anomaly detection since abnormalities in ADL are rare and data representing the anomalous cases are not readily available. The proposed Consensus Novelty Detection Ensemble approach is based on the concept of internal and external consensus. The internal consensus is an internal voting scheme within each model in the ensemble while the external consensus is an external voting scheme among the ensemble models. The weight of each model is estimated based on its performance and a score, called "Normality Score". Computed score is used in classifying the data as abnormal (anomalous) based on certain threshold crossing, normal otherwise. Experimental evaluation is conducted to detect abnormalities in ADL data obtained from CASAS repository as well as experimental dataset collected for this research. The obtained results show that the proposed approach is able to identify anomalous instances. The proposed approach offers more flexibility compared with the existing approaches by allowing the Normality Score threshold to be adjusted without retraining the models.

Keywords: Novelty Detection, Outlier Detection, Internal Consensus,

External Consensus, Abnormality Detection, Activities of Daily Living (ADL), Normality Score, Normality Threshold, Weight Estimation

1. Introduction

The ageing population, i.e. people with 65 or more years of age, is estimated be over 1.92 billion globally by 2050 [1]. It is always a preferred option for the older adults to stay in their homes [2] for as long as possible instead of looking after them in residential or care home facilities. Additionally, the cost of care for the older adults is increasing and local authorities and governments are unable to meet the financial demands [3, 4]. In order to improve their quality of life and ease the financial pressure, independent living for older adults is promoted. This will require constant monitoring of the older adults in their own homes and detecting any abnormality in their Activities of Daily Living (ADL). Abnormality is any significant deviation from individual's usual behavioural routine, and can be an early indication of Mild Cognitive Impairment (MCI) or other health-related challenges especially in ADLs that are detrimental to well-being such as sleeping, eating and toileting [5, 6].

The task of learning the behavioural routine of an individual and detecting abnormalities in it is rather arduous. Moreover, the ADL data representing human behaviour vary from one individual to another. Novelty Detection algorithms can be used to model the ADL data representing the individual daily activity routine to serve as a baseline. Subsequent activities can be compared to the baseline model to detect deviation which can be an indication of abnormality.

Novelty Detection has to do with the identification of new or unknown data. This involves detecting if a test data differs significantly from the data available during training [7]. Unlike in binary or multi-class classification where data for the different classes are available during training, in novelty detection, only one set of data is available. This is also referred to as One-Class Classification or Outlier Detection [7]. It is used for anomaly detection in a situation where there is sufficiently large training data representing the normal class and little or no training data for the anomalous (abnormal) class. This concept is relevant to ADL anomaly detection since abnormalities in ADL are rare and data representing the anomalous cases are not readily available. Novelty detection enables a model to be fitted into the

normal training data, and subsequent data to be compared to the model in order to detect abnormalities that do not conform to the built model. Some soft computing techniques based on this concept include One-Class Support Vector Machine (OC-SVM), Support Vector Data Description (SVDD), Local Outlier Factor (LOF), Isolation Forest (IF) etc. [7].

In this paper, an approach for creating an ensemble of novelty detection algorithms is proposed. The proposed Consensus Novelty Detection Ensemble (CNDE) approach generates a score for an activity termed as "Normality Score" qualifying the activity as either normal (inlier) or abnormal (outlier). Ensembles of machine learning models are usually based on a voting approach and the resulting output is a label representing the class of the data point. In the context of this research, the output can be either "normal" or "abnormal". This is not flexible for ADL anomaly detection since the threshold value for normal and abnormal activities cannot be explicitly adjusted. Human behavioural routine is complex and subject to changes due to seasonal or other factors. The "Normality Score" generated by our proposed approach allows the threshold to be dynamically adjusted to incorporate changes in the individual's routines. The proposed ensemble approach is evaluated using four heterogeneous novelty detection algorithms for the detection of ADL anomalies.

The rest of this paper is organised as follows: Section 2 details some of the related works, followed by Section 3 describing the proposed ensemble approach. Section 4 describes the dataset employed for validation of the proposed methodology and experimental results. Pertinent conclusions and future work plan are presented in Section 5.

2. Related Work

Anomaly detection in ADL has received tremendous attention over the years with different computational methodologies applied for the detection of various types of anomalies. Hoque et al. [8] proposed a system called "Holmes" for detecting ADL anomalies using Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Similarly, in [9], the number of sensor events, time and duration of an activity in a smart home is extracted and clustered with DBSCAN. Instances with unusual duration or irregular events are classified as anomalous. Jakkula et al. [10] used a method of detecting temporal relation between activities which can be classified as

anomalous. Self Organising Map (SOM) was used to detect artificially induced anomalies in occupant behaviour relating to room occupancy [11, 12].

Approaches that try to learn and recognise daily behavioural routine of an individual, then classify any deviation from the learned routine as an anomaly seems to be the most feasible approach since data representing anomalous behaviour are rarely available. Different variants of Recurrent Neural Network (RNN) such as Vanilla RNN, Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) has been applied in [5] to learn a human usual behavioural pattern and find a deviation from the learned pattern. Lotfi et al. [4] used Echo State Network (ESN) for the detection of anomalies in ADL from the raw binary sensor data. Hidden Markov Model (HMM) is trained to learn the activity sequences over a period of time and classify sequences that do not conform to what is being learned as an anomaly, then a Fuzzy Rule-Based System (FRBS) infers if the detected sequence is an actual anomaly [13].

In [14], a combination of Convolutional Neural Network (CNN) and LSTM is used to detect simulated anomalies in ADL data. Their approach is to generate synthetic anomalies mimicking the behaviour of early dementia sufferers such as disturbed sleep, repeated activities in an unknown order etc. The main dataset serves as training data for the normal class while the synthesised anomalous data serve as training data for the anomalous class. The data is then fed into a CNN in order to learn the encoding while LSTM is used to learn the activity sequences of the behavioural routine. While this approach shows a promising result, the only drawback is that the author cannot possibly generate synthetic data for each and every type of anomaly. Therefore, anomalous instances that are not generated may not be identified by the model.

Another approach to ADL anomaly detection involves clinical assessment of older adult's functional health. A health score is assigned to the older adults by an expert based on periodic evaluations in areas such as cognitive health, mobility etc. A computational model is trained to map a relationship between the ambient data collected over the period of the assessment and the assigned score. This will enable the model to predict future health score for any given data. Dawadi et al. [15] proposed Clinical Assessment using Activity Behaviour (CAAB) based on this concept and applied it to 18 smart home datasets collected for over 2 years period. Statistical correlation is established between the CAAB predicted score and the clinically assigned score. Alberdi et. al. [16] uses this same clinical assessment approach on

the same dataset utilising clinical score based on Instrumental Activities of Daily Living-Compensation (IADL-C). IADL-C consists of a larger subset of activities than CAAB such as money and self-management, home daily living, travel and event memory, and social skills. The author tries to not just predict the health score, but to also predict if there is any reliable change in the older adults activity. A regression model is used for predicting the health score while a classifier is used to predict the change. The work is extended in [17] by oversampling the minority class to cover for class imbalance. In both cases, the score predicted by the regression model shows a promising result. The classification result for predicting absolute change performs poorly, but statistical evaluation shows a correlation between the activities and the assigned clinical score. The poor performance of the classifier may be connected to the unique nature of each human behaviour.

Novelty detection algorithms have also been applied to detect anomalies in ADLs and other datasets. Authors in [18] and [19] have applied OC-SVM for ADL anomaly detection. It has also been used to detect cancerous mass in images [20]. Similarly, OC-SVM was applied for the diagnosis of faulty vehicles [21], detecting anomalies in time series data [22], applied on Electroencephalogram (EEG) data to detect seizures in human patients [23], and in combination with other novelty detectors to predicts patients that are at risk after undergoing surgery [24].

Approaches based on novelty detection that involves estimating the probability density function of the data has been proposed in which data in the region of high density are considered normal while those in the low density region are classified as anomalous [7]. The major drawback of this approach is that the data is assumed to be of certain distribution which is not practical. To overcome this, non-parametric approaches that estimate the distribution from the training data are proposed. These are applied for detection of anomalies in Jet engines [25], detection of a cancerous mass in images [26], and in the detection of network intrusion [27].

Distance measure based approaches which estimate the distance between a data point and its nearest neighbours are also studied. Data points with close neighbours are classified as inliers (normal) and those with far neighbours as outliers (anomalous). While this is seen to be computationally expensive in high dimensional space [7], it has been applied for removal of outliers in audio streams [28], and in the detection of disease outbreak [29].

Most of the proposed anomaly detection approaches in ADL are too simplistic and therefore generate a high rate of false alarm [8]. A system with

a high rate of false alarm may not be suitable for monitoring the well-being of older adults due to its unreliability. Moreover, studies have shown that a high rate of false alarm in the anomaly detection system for ADL leads to dissatisfaction by carers and clients [30]. To restrict the false alarm, the behaviour of the user needs to be modelled accurately. This can be achieved by using an ensemble of novelty detection algorithms since each model is good on certain characteristic features. For example, OC-SVM is sensitive to the presence of outliers in the training data thereby resulting in poor performance while IF is able to perform well even when the training data is contaminated with outliers because it isolates the anomalies instead of profiling the normal data [31][32]. An ensemble of machine learning models combines multiple model's predictions to achieve better accuracy. For anomaly detection, an ensemble of homogeneous algorithms that produces the same output is not good enough. However, an ensemble of heterogeneous algorithms is better since it will provide the much needed diversity and accuracy [33].

In [33], an approach for creating an ensemble of outlier detectors using similarity measures is proposed while Dib et al. [34] applied an ensemble of novelty detection models for damage detection for structural health monitoring.

The ensemble of novelty detection models has not been given much attention and according to our knowledge, none of the few proposed approaches takes the concept of Normality Score into consideration. The Normality Score approach gives more flexibility since the threshold for the score signifying inliers and outliers can be adjusted dynamically. Table 1 summarises some of the related research works in the area of ADL anomaly detection. Due to the utilisation of different datasets by the different authors as well as variability in the evaluation criteria, only the data collection and sensing modality, computational methodology, nature of the experiment, and their evaluation metrics are highlighted.

Table 1: A summary of related research studies in ADL anomaly detection (Abbr.: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Event Error Rate (EER))

Ref.	**Data and Sensor**	**Approach**	**Scope of Work and Limitations**	**Evaluation Metrics**
[10]	Inferred ADL data	Temporal relation of activities	Detecting synthetic anomalies. The approach relies on activities with temporal relation. Anomalies in non-temporal activities may not be identified.	Percentage of correct prediction
[15]	Inferred ADL data	CAAB	Predicting health score. The physiological state of the subject during the assessment may lead to wrong score assignment in which this approach relies on.	Correlation and RMSE
[19]	Inferred ADL data	OC-SVM	Detecting synthetic anomalies. While the obtained result is promising, comparison with other anomaly detection models is not carried out	Accuracy
[17] [16]	Inferred ADL data	IADL-C	Predicting health score, Detecting health decline. The score predictor performs well while model for detecting health decline achieved low accuracy.	RMSE, Correlation, F-Score, Accuracy, Sensitivity and ROC-AUC
[14]	Raw binary data	CNN + LSTM	Detecting synthetic anomalies. Non simulated anomalies may not be identified.	Sensitivity, Specificity
[4]	Raw binary data & Inferred ADL data	ESN	Detecting synthetic anomalies. Data for only one subject is used. Data from multiple subjects is required to test for generalisation.	RMSE
[8]	Inferred ADL data	DBSCAN	Monitoring and Detecting anomalies	Precision and Recall
[18]	Inferred ADL data	OC-SVM	Detect artificially induced anomalies	Error Rate, F-Measure and Accuracy
[5]	Raw binary data	Vanilla RNN, GRU, LSTM	Detecting synthetic anomalies. This approach may not detect anomalies that are not simulated	Accuracy
[9]	Inferred ADL data	DBSCAN	Detecting anomalies in data	EER
[13]	Inferred ADL data	HMM + FRBS	Detecting ADL anomalies, Detection early disease symptoms	Accuracy, Precision, Recall, F-Measure, ROC-AUC
[11] [12]	Raw binary data	SOM	Detecting both real and synthetic anomalies	Accuracy

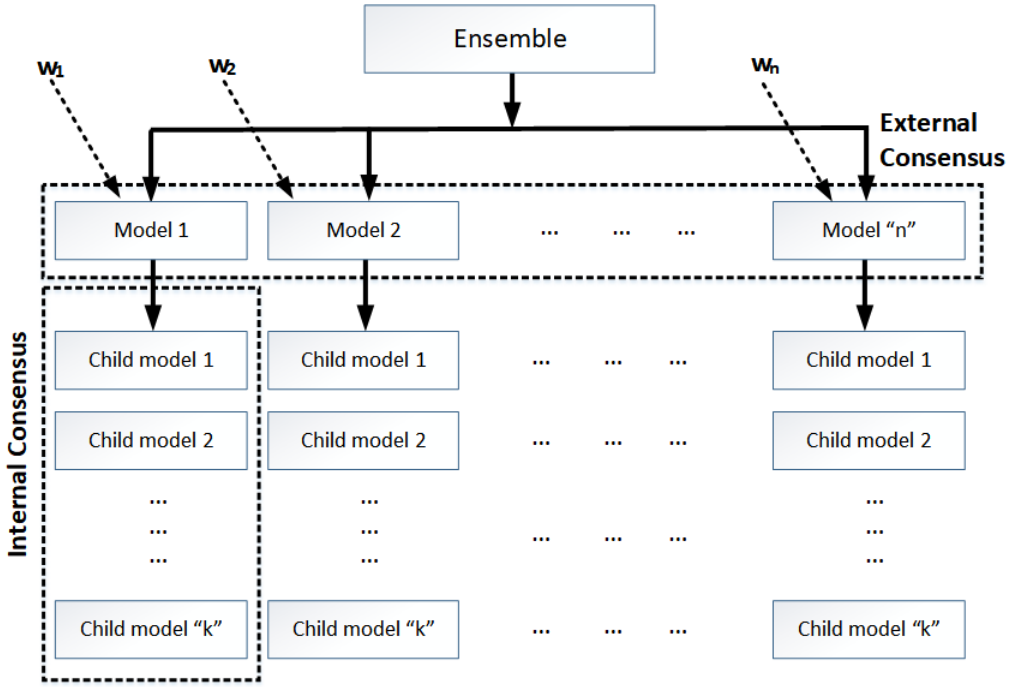


Figure 1: An schematic diagram of the proposed CNDE.

3. Methodology

The proposed CNDE approach is based on the concept of internal and external consensus. This is inspired by the concept proposed by Mahmud et al. [35] in which a sensor node is certified based on its data and behavioural trust among the other nodes in the infrastructure. The internal consensus is an internal voting scheme within each model in the ensemble (i.e. a number of child models is created for each model and their votes are aggregated and a score is computed for the data points). The external consensus is a voting scheme among the models in the ensemble similar to majority vote approach. Appropriate weights are estimated and assigned to the respective models in the ensemble based on their performance. The Normality Score generated by the CNDE enables the data to be classified as either normal or abnormal. A higher Normality Score indicates that the data point is an inlier (normal) while a lower Normality Score signifies an outlier (anomaly). Figure 1 illustrates a schematic diagram of the CNDE approach.

Given a set of n training samples $X = \{x_1, x_2, ..., x_n\}$ of $d-$dimensional

data (i.e., $X \in \mathbb{R}^d$), let $A = \{a_1, a_2, ..., a_m\}$ be a set of m models. A data point x can be classified as either an outlier or inlier by computing its Normality Score $\mathcal{N}_x$ using an ensemble of A as expressed in Equation 1 below.

$$\mathcal{N}_x = \frac{F_x^c + E_x^c}{2} \tag{1}$$

F_x^c is the Combined Internal Consensus Score (CICS) and E_x^c is the Combined External Consensus Score (CECS) of the data point x respectively.

$$E_x^c = \frac{1}{m} \sum v_x \tag{2}$$

v_x is the number of votes x received as an inlier from the models and m is the number of models in the ensemble.

The aggregate of the votes v_x is termed as the Combined External Consensus Vote (CECV). The class of a data point based on external consensus is determined by the majority vote of the CECV.

The CICS is the weighted average of the Internal Consensus Score (ICS) as expressed in Equation 3.

$$F_x^c = \frac{1}{m} \sum_{i=1}^{m} I_x^i * w_i \tag{3}$$

where m is the number of models in the ensemble, I_x^i is ICS of the i^{th} model for the data point x, and w_i is the weight of the i^{th} model.

The ICS expressed in Equation 4 is inspired by Bagging approach in machine learning [36]. The training data is split randomly into k-folds. A k-child models are created for each model in the ensemble. The k-child models are trained each with one separate fold out of the k-fold training data as illustrated in Figure 1. The votes a data point x receives from the k-child models are termed as the Internal Consensus Vote (ICV). A data point is considered an inlier by a model if it has 1 or more ICV.

$$I_x = \frac{1}{k} \sum_{i=1}^{k} v \tag{4}$$

v is the number of votes x received as an inlier from the child models, and k is the number of child models (i.e. number of folds).

The difference between the ICS and the CICS is, the ICS determines the

score of a data point for an individual model in the ensemble while CICS computes the score of the data point across all the models in the ensemble using the respective models' weights.

The weight of each model is a value ranging from 0 to 1. The models performing better receives larger weight and vice versa. This is estimated during training since it is impossible to manually assign appropriate weight values. The weight of each model is initialised to 1 and penalised by the percentage of wrong predictions made by the model. Wrong predictions are determined by comparing the CECV and the ICV. Variability of prediction between the CECV and ICV is considered a wrong prediction, therefore, warrants for the penalisation of that model.

Let $W = \{w_1, w_2, ..., w_m\}$ be the weights of the m models in the ensemble, the final weight of each individual model after penalisation can be expressed as:

$$w_f = w_i - \frac{e}{n} w_i \tag{5}$$

where w_f is the final weight after penalisation, w_i is the initial weight before penalisation (i.e. 1), e is the number of wrong predictions made by the i^{th} model and n is size of the training samples.

Introducing a threshold value ε termed as "Normality Threshold" to serve as cut-off point for the Normality Score, the function $f(x)$ that determines the class of x is expressed as:

$$f(x) = \begin{cases} \mathcal{N}_x \geq \varepsilon & then \quad x \quad is \quad an \quad Inlier \\ \\ \mathcal{N}_x < \varepsilon & then \quad x_i \quad is \quad an \quad Outlier \end{cases} \tag{6}$$

The Normality Score is a value ranging from 0 to 1 (i.e. $0 \leq \mathcal{N}_x \leq 1$) with higher score signifying inlier and lower score signifying outlier. Certain standard deviations to the left of the Normality Score is considered as the threshold as shown in Figure 2. An ideal threshold is -3σ (i.e. any score below -3σ is an outlier).

Algorithm 1, Algorithm 2, Algorithm 3 and Algorithm 4 shows the procedures for computing the CECS, CICS, ICS and Weights of the models respectively as described above. The time complexities of the respective algorithms are tabulated in Table 2. The worst-case complexities (O) of the algorithms for the Weight Estimation and Combined Internal Consensus are dependent not just on the size of the input data (n) but also on the num-

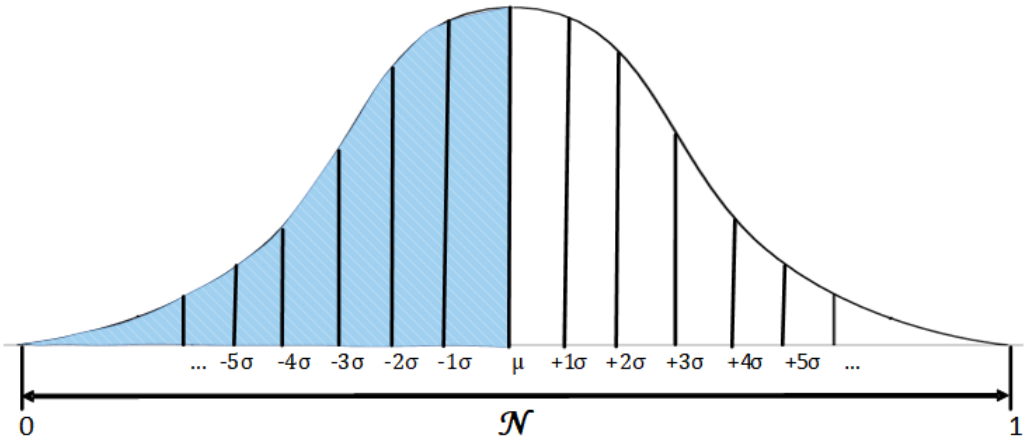


Figure 2: Normality threshold estimation.

ber of models in the ensemble (m). It is highly unlikely that the number of models employed will exceed the size of the training data, and therefore, the complexity will most likely be of the form $O(n)$. However, if in any case, the number of models in the ensemble is equal to the size of the training data, then the complexity will be $O(n^2)$. An improved time complexity can be achieved by maintaining a fixed number of models in the ensemble and having the size of the training data larger than the number of models.

3.1. Combined Concepts

Combining the concepts described, the diagram in Figure 3 shows the training and testing phases of the proposed approach. The training phase involves the random split of the training data into k-folds, creation of k-child models for each individual model in the ensemble, training and estimating an appropriate weight for each model, computing the Normality Score of the training data and estimating an optimal Normality Threshold. During the

Table 2: Time complexities of proposed algorithms.

Algorithm	Complexity
Combined External Consensus Algorithm (CECA)	$O(n)$
Combined Internal Consensus Algorithm (CICA)	$O(n * m)$
Internal Consensus Algorithm (ICA)	$O(n)$
Weight Estimation Algorithm(WEA)	$O(n * m)$

Algorithm 1 Combined External Consensus Algorithm (CECA)

Input: Dataset $X = \{x_1, x_2, ..., x_n\}$,
Models List $A = \{a_1, a_2, ..., a_m\}$
Output: Vote and Score $(V^e, E^c) = \{(v_{x_1}, e_{x_1}), (v_{x_2}, e_{x_2}), ..., (v_{x_n}, e_{x_n})\}$

procedure CECA
  for each $x \in X$ **do**
    $t = \sum v$ ▷ Aggregate votes of x as inlier from A
    $e_x = \frac{1}{m}t$ ▷ Computing the CECS of x
    if t is the majority **then**
      $v_x = 1$ ▷ x is an inlier by CECV
    else
      $v_x = 0$ ▷ x is an outlier by CECV
    end if
    $(V^e, E^c) \leftarrow (v_x, e_x)$ ▷ Append result to (V^e, E^c)
  end for
  return (V^e, E^c) ▷ Return the CECV and CECS of x

testing phase, the trained models along with their estimated weights and the Normality Threshold are applied to predict the class of the test data (i.e. either inliers or outliers).

4. Experimental Results

In this section, the proposed CNDE model in Section 3 is evaluated. Datasets used for the validation, extracted features and obtained results are also described.

4.1. Data Description

Two separate data sets representing the ADL of older adults are employed for the validation of the proposed methodology. More details about the datasets are provided in the following sections.

4.1.1. Activities of Daily Living Dataset

Data is collected for a single resident for a period of 72 days. Low-cost non-intrusive ambient sensors such as Passive Infrared (PIR), Pressure and Door sensor are used as the data collection devices. This data collection modality is the most widely accepted method for ADL monitoring due to its non-invasive nature as compared to vision-based approach (e.g. using cameras) which studies have shown that it is widely rejected due to privacy and ethical concerns [37].

Algorithm 2 Combined Internal Consensus Algorithm (CICA)

Input: Dataset $X = \{x_1, x_2, ..., x_n\}$,
Models List $A = \{a_1, a_2, ..., a_m\}$,
Models Weights $W = \{w_1, w_2, ..., w_m\}$
Output: CICS $F^c = \{f^c_{x_1}, f^c_{x_2}, ..., f^c_{x_n}\}$

procedure CICA
  for each $x \in X$ **do**
    for each $a \in A$ **do**
      $I_x =$ Compute the ICS of x
      $\rho \leftarrow I_x * w_a$ ▷ ICS and model's weight
    end for
    $f^c_x = \frac{1}{m} \sum_{i=1}^{m} \rho$ ▷ Compute CICS for x
    $F^c \leftarrow f^c_x$ ▷ Append result to F^c
  end for
  return F^c ▷ Return the CICS of x

Data generated by these sensors are binary in nature with 1 and 0 signifying active and inactive states respectively. Activities performed by the residents are inferred from the sensor readings. For example, the firing of the PIR sensor in the restroom signifies that the resident is using the restroom while that of the pressure sensor on the bed is an indication that the resident is sleeping. Figure 4 shows a pictorial representation of the inferred activities from the binary sensors' data.

Activities recorded include preparing a meal (kitchen activity), eating (dining room activity), staying in the living room, toileting, going out of the house, and sleeping. Each activity has its start time, an end time, and in some cases, the location of the performed activity as shown in Table 3.

Table 3: Sample of collected ADL data.

Activity	Start Time	End Time
Dining Room	2018-05-01 17:19:31	2018-05-01 17:28:45
Living Room	2018-05-01 17:28:59	2018-05-01 20:34:31
Toilet	2018-05-01 20:34:41	2018-05-01 20:42:07
Bedroom - Sleeping	2018-05-01 22:49:43	2018-05-02 07:46:07
...	...	...

Algorithm 3 Internal Consensus Algorithm (ICA)

Input: Dataset $X = \{x_1, x_2, ..., x_n\}$,
Model M with k-child models $M = \{m_1, m_2, ..., m_k\}$
Output: Vote and Score $(V^i, I) = \{(v_{x_1}, i_{x_1}), (v_{x_2}, i_{x_2}), ..., (v_{x_n}, i_{x_n})\}$

procedure ICA
  for each $x \in X$ **do**
    $t = \sum v$ ▷ Aggregate votes of x as inlier from childrens of M
    $i_x = \frac{1}{k}t$ ▷ Computing the ICS of x
    if $t \geq 1$ **then**
      $v_x = 1$ ▷ x is an inlier
    else
      $v_x = 0$ ▷ x is an outlier
    end if
    $(V^i, I) \leftarrow (v_x, i_x)$ ▷ Append result to (V^i, I)
  end for
  return (V^i, I) ▷ Return the ICV and ICS of x

Algorithm 4 Weight Estimation Algorithm (WEA)

Input: Dataset $X = \{x_1, x_2, ..., x_n\}$,
Models List $A = \{a_1, a_2, ..., a_m\}$
Output: Weights $W = \{w_1, w_2, ..., w_m\}$

procedure WEA
  Initialise weights $W = \{w_1, w_2, ..., w_m\}$ to 1
  Initialise errors $E = \{e_1, e_2, e_m\}$ to 0
  for each $x \in X$ **do**
    v_e = Get the CECV of x
    for each $a \in A$ **do**
      v_i = Get the ICV of x by model a
      if $v_e \neq v_i$ **then**
        $e_a = e_a + 1$ ▷ Increment error count of model a
      end if
    end for
  end for
  for each $a \in A$ **do**
    w_a = Get the weight of model a from W
    e_a = Get the errors of model a from E
    $w_a^* = w_a - \frac{e_a}{n} w_a$ ▷ Compute final weight by penalisation
    $W \leftarrow w_a^*$ ▷ Update weight of model a
  end for
  return W ▷ Return estimated weights of A

4.1.2. CASAS H111 Dataset

The H111 dataset from CASAS repository [38] is also used to evaluate the proposed algorithm. It also provides a benchmark to compare the per-

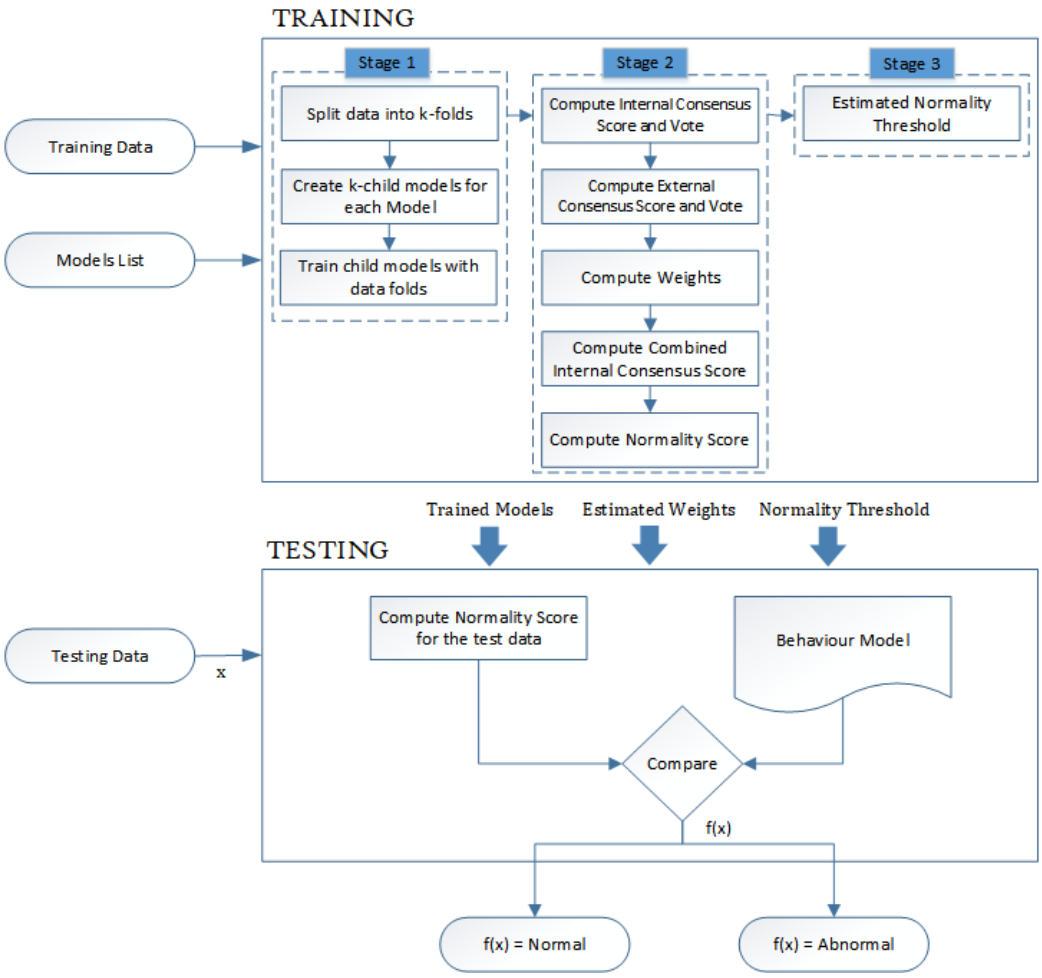


Figure 3: The training and testing phases of the proposed CNDE approach.

formance of the proposed algorithm. The dataset contains the ADL of a volunteer adult living alone in his residence for a period of 50 days. Activities recorded include sleeping, eating, bathing, dressing toileting etc. The dataset does not provide any information as to whether there is an abnormality in the resident's activity or not.

Our approach involves training the model with data for a certain number of days (e.g. 31 days) and test it against data of the remaining days. Different anomalous cases may also be simulated such as going to bed early/late, oversleeping etc. to certify the model's ability in predicting these behaviours.

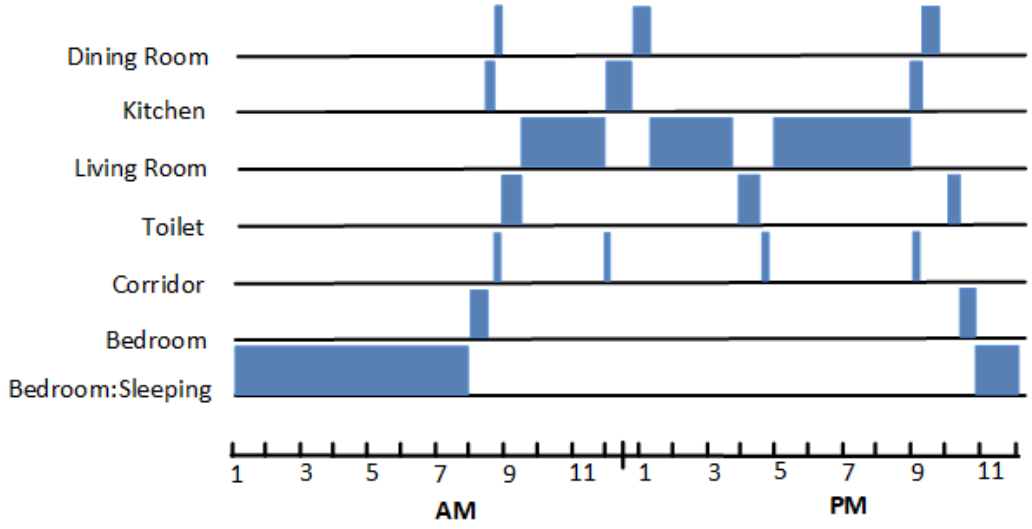


Figure 4: A sample of combined activities inferred from ambient sensors.

4.2. Data Pre-processing

The datasets described above contains different ADLs. For this experiment, only sleeping activity is filtered and selected. Relevant features that can discriminate between the normal and anomalous cases are selected.

- Start time: This is the starting hour and minutes of the activity. The start hour ranges from 0 to 23. It is then converted to a scale of -11 to +11 with 0 representing 12 midnight. This is because generally, people do go to bed at night time. An activity that starts at 11:50 pm is closer to that of 1:00 am in terms of the start time than an activity performed at 9:00 pm. However, without converting the start time to a scale of -11 to $+11$, the margin between 11:50 pm to 1:00 am will be larger as shown in Figure 5.

- Duration: This is the duration in minutes of the activity obtained by subtracting the start time from the end time.

- Number of interruption: This is the number of times an individual leaves the bed and returns back to it. For example, an individual may leave the bed in the middle of the night to use the restroom. If the interval (in minutes) between the time the individual leaves the bed

and returns to it is less than an hour (60 minutes), it is considered an interruption, else it is assumed that the activity has ended.

- Duration of interruption: This is the total duration (in minutes) of all the interruptions within an activity.

- Day of activity: This represents the day in which the activity is performed ranging from 0 to 6 representing the 7 days of a week. This is important because the individual may go to bed late some days and early some other days due to his/her routine e.g. watching a specific late-night TV show every Monday.

- Weekend or Weekday: This is to determines if the activity is performed on weekdays or weekends. Some individuals might have a different routine for weekdays and weekends while some might not. 0 and 1 represents weekdays and weekends respectively.

Because the extracted features are in different scales, models sensitive to scaling may perform poorly on the dataset. The selected features are normalised.

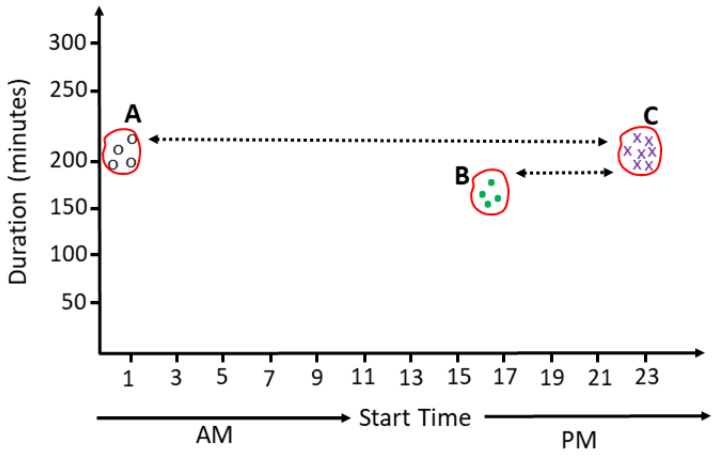
4.3. Model Selection and Optimal Parameters

The proposed CNDE approach described in Section 3 is generic and can be used with any number of novelty detection models. For this empirical evaluation, models employed are Isolation Forest (IF), One-Class SVM (OC-SVM) with Radial Basis Function (RBF) kernel, Local Outlier Factor (LOF), and Robust Covariance Estimation (RCE). The contamination rate (i.e. the rate of outliers in the training data) is set to 0.1 across all the models since the aim is to model the training data with minimal error.

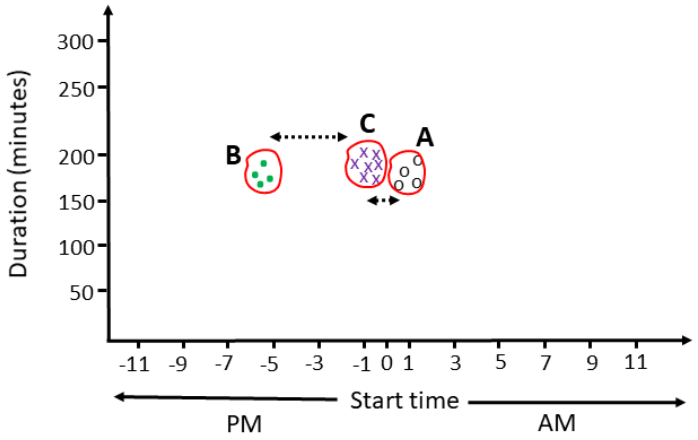
The number of folds for the Internal Consensus can be varied depending on the size of the training set. Three (3) folds are used for both our collected data and CASAS H111 dataset, with 31 days data used for training. The weights of the respective models are initialised to 1 and the Normality Threshold is taken as -3σ as described in the previous sections.

4.4. Ensemble Approach Evaluation

This section contains the obtained results for both the collected and CASAS H111 dataset.



(a)

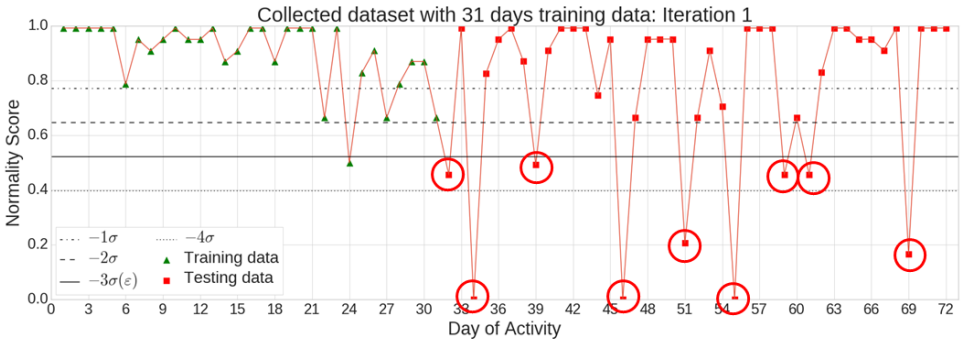


(b)

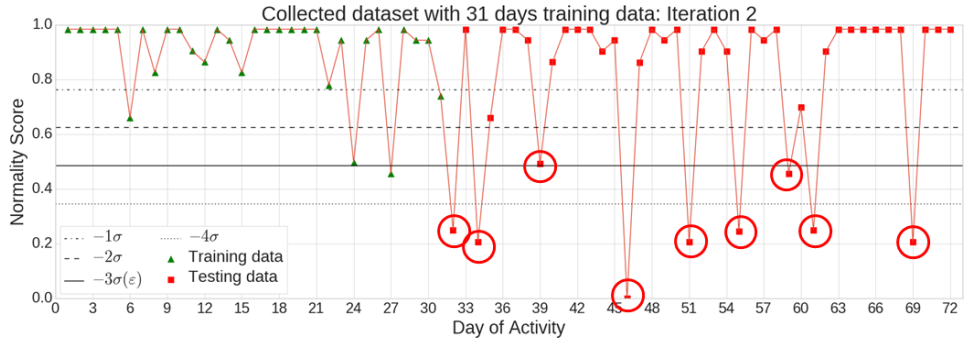
Figure 5: Clusters with default and converted start time; a) default start time, b) converted start time.

4.4.1. Activities of Daily Living Dataset

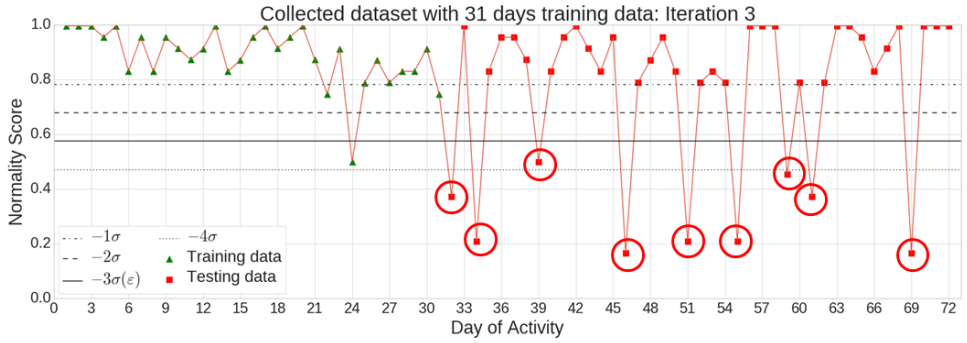
An experiment has been conducted on the described collected data. The first 31 days data is used to train the model while the remaining 41 days data is used for testing. To ensure that the proposed ensemble approach generalises, multiple iterations of the experiment is run since each iteration



(a)



(b)



(c)

Figure 6: Normality Score for the collected data; a) Iteration 1, b) Iteration 2, c) Iteration 3

splits and shuffles the data randomly.

The result in Figure 6 shows a plot of the Normality Score for the collected data for 3 iterations. Even though the Normality Score varies across the iterations, the difference is negligible. The Nine (9) days identified as anomalous are the same across all the iterations. The dataset is examined for variations between the days identified as anomalous and those identified as normal. Table 4 summarises the findings:

It can be seen that the model identifies data points that do not conform to the known individual's behavioural routine even though it miss-classified 2 days as anomalous (i.e. Day 32 and Day 61).

A test is conducted with 18 days of training data to verify if the size of the training data has any significant effect on the model's performance. The obtained result is shown in Figure 7. It can be seen that the model performs poorly when trained with data for 18 days as compared to 31 days of training data.

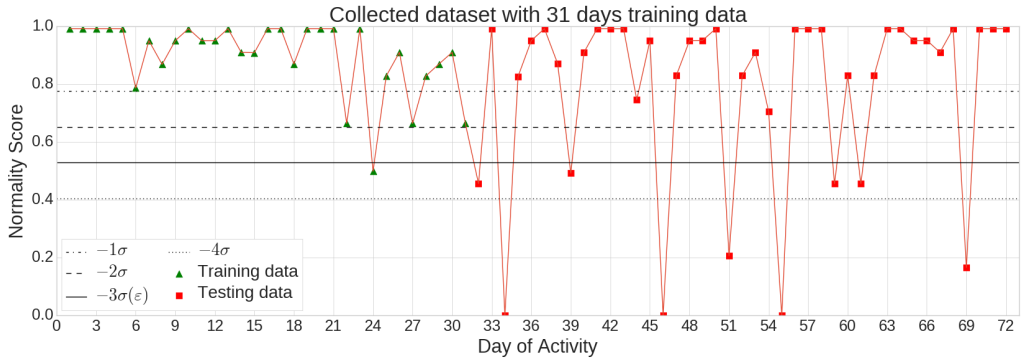
4.4.2. CASAS H111 Dataset

A test is conducted on the CASAS H111 data with 31 days of data used for training and the remaining data for testing. Multiple iterations are run to verify generalisation.

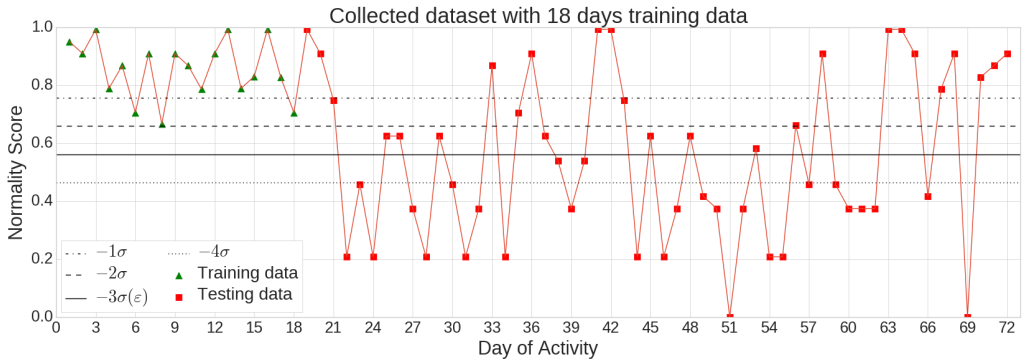
The Normality Score plot for the CASAS H111 dataset for 3 different iterations is shown in Figure 8. The identified anomalous days are examined and the findings are summarised in Table 5.

Table 4: A summary of identified anomalies and possible causes for the collected ADL data.

Day	**Cause**	**Detailed Description**
Day 34, 46 & 55	Less Sleeping	The participant sleep for short period of time compared to the usual duration
Day 51	Over Sleeping	The participant sleep for a longer duration than the usual.
Day 39 & 69	Interrupted Sleep	The individual has multiple transitions from bed to other locations
Day 59	Late Sleep	The individual goes to bed late
Day 32 & 61	Model Error	No deviation has been identified from the usual routine



(a)



(b)

Figure 7: Normality Score for 31 days and 18 days training data (collected data); a) 31 days training data, b) 18 days training data.

The model is able to detect the data points that do not conform to the known resident's behavioural routine with the exception of Day 36 which is miss-classified. Similarly, a test is conducted with 18 days of training data and the result is shown in Figure 9.

From Figure 9, it can be seen that the results of both 18 days and 31 days training data are comparably similar unlike in the case of the collected data. This further proves our initial assertion that human behavioural routine varies from one individual to the other. However, it can be established that the minimum number of days required for modelling ADL behavioural routine is 31 days.

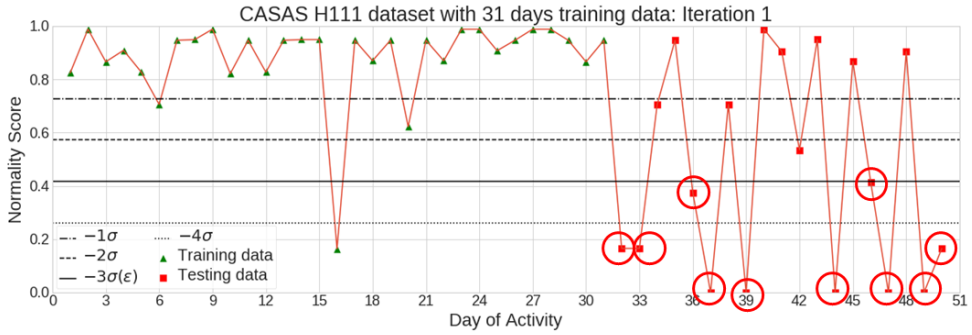
4.5. Comparison with Ensemble Methods

To evaluate the proposed ensemble approach, a comparison is made with ensemble approach based on majority vote as well as approaches proposed in [33], namely; Ensemble of Detectors with Correlated Votes (EDCV) and Ensemble of Detectors with Variability Votes (EDVV). The ensemble approach based on majority vote involves the respective models in the ensemble voting the data as either inliers or outliers with the class having the majority votes taken as the final prediction. Both 2 and 3 are used as the value of the majority vote threshold since the ensemble contains only 4 models. The EDCV and EDVV are similar to the majority vote approach except that the for the EDCV, weights of the ensemble model is estimated from the correlation coefficient of the models' predicted score, while for the EDVV, the weights are estimated from the Mean Absolute Deviation (MAD) of the prediction score. More details on these approaches can be found in [33].

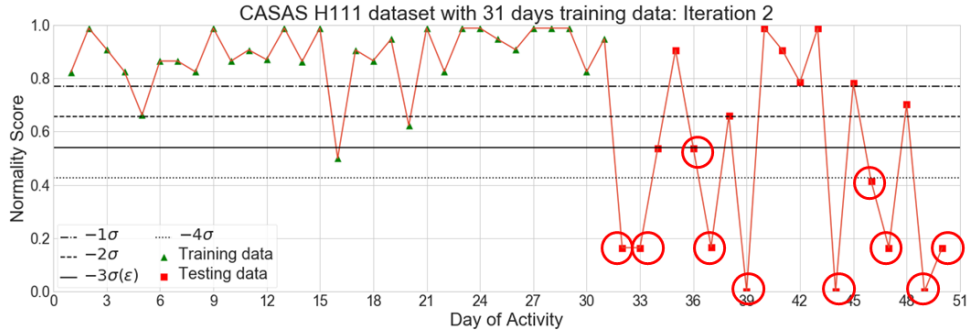
In order to measure the performance metrics, synthetic anomalous data is generated for a period of 100 days. The synthetic data is generated to

Table 5: Summary of identified anomalies and possible causes for the CASAS H111 data.

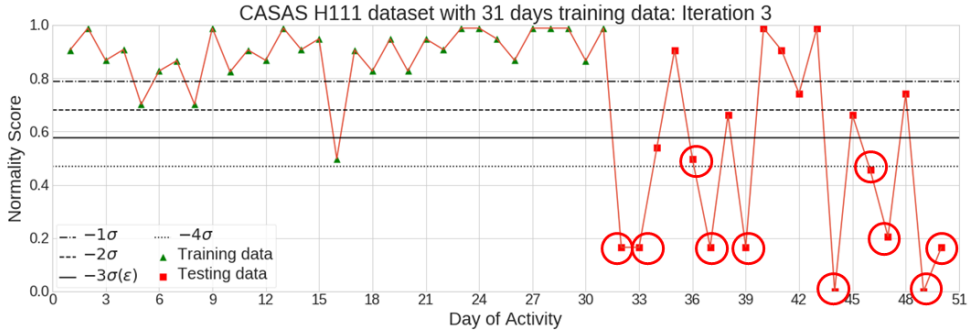
Day	**Cause**	**Detailed Description**
Day 32, 33 & 47	Afternoon Sleep	The resident had over 1-2 hours nap during the day. There is never an instance where the resident sleeps during the day in the data used for training
Day 39 & 49	Interrupted Sleep	The resident has multiple transition from bed to other locations. The number of transition is twice what the model has seen during training
Day 37, 44, 46	Longer Interruption	The duration of interruption the resident had is longer than the usual
Day 50	Less Sleeping	The individual only sleeps for approximately 2 hours
Day 36	Model Error	No significant variation is identified from the usual routine



(a)

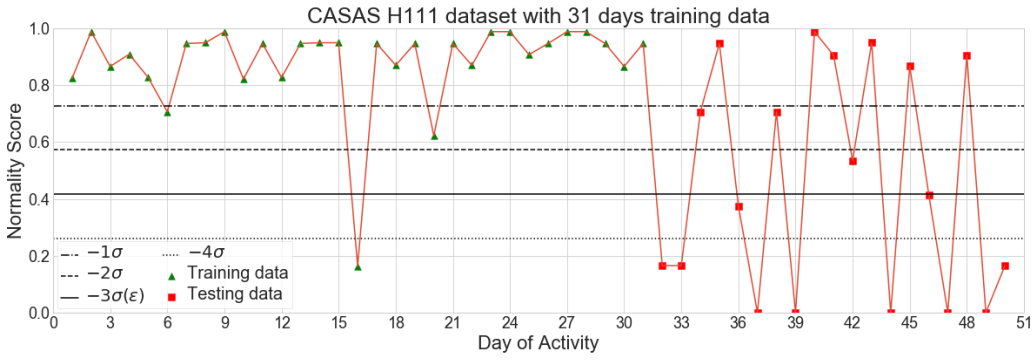


(b)

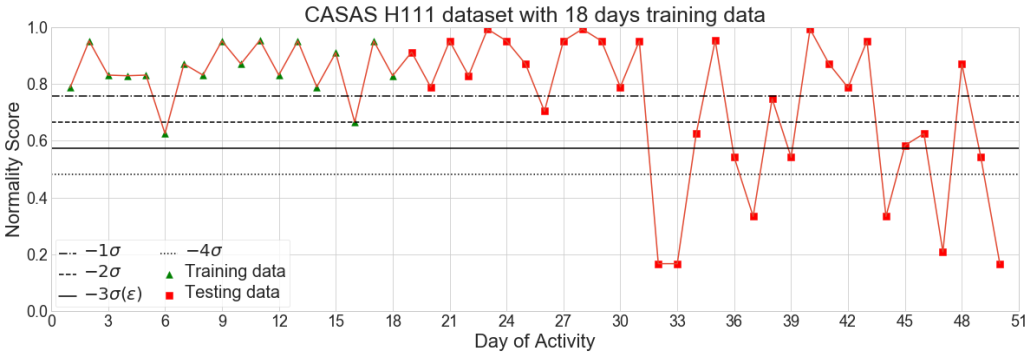


(c)

Figure 8: Normality Score for CASAS H111 dataset; a) Iteration 1, b) Iteration 2, c) Iteration 3.



(a)



(b)

Figure 9: Normality Score for 31 days and 18 days training data (CASAS H111 data); a) 31 days training data, b) 18 days training data.

simulate different anomalous instances such as going to bed late, insufficient sleep, interrupted sleep etc. The training data (both collected and H111 dataset) are oversampled using Synthetic Minority Over-sampling Technique (SMOTE) proposed in [39] so that the problem of class imbalance is eliminated. The same models (i.e. OC-SVM, IF, LOF and RCE) are utilised in the various ensemble approaches. The results obtained from the comparison is presented in Table 6.

Considering the presented results, the proposed approach achieved a better performance than the other approaches. The only exception is in the case of our collected data when 18 days data is used to train the model. This is not surprising since it is established earlier that 31 days of training data is the minimum required for behaviour modelling. Similar to what is obtained

in the Normality Score plot in Figure 9, The results for CASAS H111 dataset when trained with data for 18 days and 31 days are comparably similar, while for the collected data, the results are significantly different. This is another confirmation of the variability in the behavioural routine of one individual to another. Overall, the ensemble approach based on majority votes with a 3 votes threshold outperformed that of a 2 votes threshold.

4.6. Discussion

Based of the results presented in Table 6, it can be argued that the proposed method outperformed other ensemble approaches. It can also be observed that the proposed approach has nearly a linear time complexity in all cases.

The weight estimation algorithm allows for easy identification of better performing model for any given dataset. This is important since poor performing models for a given dataset can be identified and removed from the ensemble. The Normality threshold can be adjusted without explicitly retraining the models giving more flexibility to incorporate changes in the overly changing human behavioural routine.

Novelty detection models are created on the promise that there is only one set of available training data. This means that the training data contains none or a negligible amount of outliers [7]. A significant amount of

Table 6: Result of comparison with other ensemble methods based on accuracy.

	Training Data (Days = 31)		**Training Data (Days = 18)**	
Ensemble Approach	**Collected Data**	**CASAS H111**	**Collected Data**	**CASAS H111**
EDCV	0.92958	0.83099	0.88732	0.85915
EDVV	0.90141	0.81690	**0.91549**	0.83099
Majority Vote (v=2)	0.54930	0.77465	0.54930	0.78873
Majority Vote (v=3)	0.76056	0.92958	0.60563	0.94366
Our Approach (CNDE)	**0.98592**	**0.95775**	0.77465	**0.97183**

outliers (noise) in the training data can drastically affect the performance of the models, and therefore, the proposed ensemble approach. To address this problem, the most feasible approach is to reduce the class imbalance problem by undersampling the majority class or oversampling the minority class. Supervised learning algorithms can then be utilised to classify the data as applied in [17].

While the proposed ensemble approach is applied in a batch manner where all the needed training data are available, it has the potential of being utilised in an online or incremental learning scenario where the models are required to adapt to new data as they become available. This is possible because the threshold value qualifying the data as inliers or outliers can be adjusted without explicitly retraining the models. The distribution of the normality score for the new incoming data generated by the ensemble approach can be used to estimate a new threshold value.

5. Conclusion

In this paper, an ensemble approach for novelty detection algorithms is proposed based on the concept of internal and external consensus. The proposed CNDE approach is applied for detection of ADL anomalies. Experiments conducted on both collected ADL data and H111 data obtained from CASAS repository produced an excellent result.

In addition, the weights of the models in the ensemble are estimated during training based on the models' performance allowing for the identification of suitable models to be included in the ensemble. The resulting output of the ensemble approach is a score termed as "Normality Score" qualifying the data as inliers or outliers. Due to the dynamic nature of human behavioural routine, the proposed approach offers more flexibility since the threshold of the Normality Score can be dynamically adjusted to incorporate changes in human activities. The dynamic threshold enables new or unknown activities to be incorporated into the anomaly detection model in an incremental manner without retraining the entire ensemble models.

Further work in this area will include testing the proposed approach on longitudinal ADL data to determine long term behavioural changes and perform trend analysis. Performance metrics will also be evaluated by experimenting on a large labelled anomalous data. While the proposed approach is able to detect ADL anomalies, features of the dataset that are likely to

be the cause of the anomaly cannot be identified. This will be addressed as part of the upcoming future works.

Conflict of interest statement

There is no conflict of interest.

Acknowledgements

This research project is supported by Nottingham Trent University through Vice Chancellor Studentship Scheme provided to Salisu Wada Yahaya.

References

[1] S. Chernbumroong, S. Cang, A. Atkins, H. Yu, Elderly activities recognition and classification for applications in assisted living, Expert Systems with Applications 40 (5) (2013) 1662 – 1674.

[2] P. Rashidi, A. Mihailidis, A survey on ambient-assisted living tools for older adults, IEEE Journal of Biomedical and Health Informatics 17 (3) (2013) 579–590.

[3] Q. Ni, A. B. García Hernando, I. P. De la Cruz, The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development, Sensors 15 (5) (2015) 11312–11362.

[4] A. Lotfi, C. Langensiepen, S. M. Mahmoud, M. J. Akhlaghinia, Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour, Journal of Ambient Intelligence and Humanized Computing 3 (3) (2012) 205–218.

[5] D. Arifoglu, A. Bouchachia, Activity recognition and abnormal behaviour detection with recurrent neural networks, Procedia Computer Science 110 (2017) 86 – 93.

[6] M. Borazio, E. Berlin, N. Kücükyildiz, P. Scholl, K. V. Laerhoven, Towards benchmarked sleep detection with wrist-worn sensing units, in: 2014 IEEE International Conference on Healthcare Informatics, 2014, pp. 125–134.

[7] M. A. Pimentel, D. A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Processing 99 (2014) 215 – 249.

[8] E. Hoque, R. F. Dickerson, S. M. Preum, M. Hanson, A. Barth, J. A. Stankovic, Holmes: A comprehensive anomaly detection system for daily in-home activities, in: International Conference on Distributed Computing in Sensor Systems, IEEE, Fortaleza, Brazil, 2015, pp. 40–51.

[9] L. G. Fahad, M. Rajarajan, Anomalies detection in smart-home activities, in: The 14th International Conference on Machine Learning and Applications (ICMLA), IEEE, Miami, USA, 2015, pp. 419–422.

[10] V. Jakkula, D. J. Cook, A. S. Crandall, Temporal pattern discovery for anomaly detection in a smart home, in: The 3rd IET International Conference on Intelligent Environments, IET, Ulm, Germany, 2007, pp. 339–345.

[11] M. Novák, M. Biňas, F. Jakab, Unobtrusive anomaly detection in presence of elderly in a smart-home environment, in: 2012 ELEKTRO, 2012, pp. 341–344.

[12] M. Novák, F. Jakab, L. Lain, Anomaly detection in user daily patterns in smart-home environment, Journal of Selected Areas in Health Informatics (JSHI) 3 (6) (2013) 1–11.

[13] A. R. M. Forkan, I. Khalil, Z. Tari, S. Foufou, A. Bouras, A context-aware approach for long-term behavioural change detection and abnormality prediction in ambient assisted living, Pattern Recognition 48 (3) (2015) 628 – 641.

[14] D. Arifoglu, A. Bouchachia, Detection of abnormal behaviour for dementia sufferers using Convolutional Neural Networks, Artificial Intelligence in Medicine 94 (2019) 88–95.

[15] P. N. Dawadi, D. J. Cook, M. Schmitter-Edgecombe, Automated Cognitive Health Assessment from Smart Home-Based Behavior Data, IEEE Journal of Biomedical and Health Informatics 20 (4) (2016) 1188–1194.

[16] A. A. Aramendi, A. Weakley, A. A. Goenaga, M. Schmitter-Edgecombe, D. J. Cook, Automatic assessment of functional health decline in older

adults based on smart home data, Journal of Biomedical Informatics 81 (2018) 119 – 130.

[17] A. Alberdi, A. Weakley, M. Schmitter-Edgecombe, D. J. Cook, A. Aztiria, A. Basarab, M. Barrenechea, Smart home-based prediction of multidomain symptoms related to alzheimer's disease, IEEE Journal of Biomedical and Health Informatics 22 (6) (2018) 1720–1731.

[18] V. Jakkula, D. J. Cook, Detecting anomalous sensor events in smart home data for enhancing the living experience, in: Proceedings of the 7th AAAI Conference on Artificial Intelligence and Smarter Living: The Conquest of Complexity, AAAI Press, 2011, pp. 33–37.

[19] S. W. Yahaya, C. Langensiepen, A. Lotfi, Anomaly detection in activities of daily living using one-class support vector machine, in: Advances in Computational Intelligence Systems, Springer, 2019, pp. 362–371.

[20] S. Dreiseitl, M. Osl, C. Scheibb"ock, M. Binder, Outlier Detection with One-Class SVMs: An Application to Melanoma Prognosis, Proceedings of the AMIA Annual Fall Symposium.

[21] A. Theissler, Multi-class Novelty Detection in Diagnostic Trouble Codes from Repair Shops, in: 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), 2017, pp. 750–763.

[22] J. Ma, S. Perkins, Time-series novelty detection using one-class support vector machines, Proceedings of the International Joint Conference on Neural Networks.

[23] A. Gardner, A. Krieger, G. Vachtsevanos, B. Litt, One-class novelty detection for seizure analysis from intracranial EEG, Journal of Machine Learning Research.

[24] Z. Syed, M. Saeed, I. Rubinfeld, Identifying High-Risk Patients without Labeled Training Data: Anomaly Detection Methodologies to Predict Adverse Outcomes, AMIA Annual Symposium proceedings.

[25] A. Nairac, T. A. Corbett-Clark, R. Ripley, N. W. Townsend, L. Tarassenko, Choosing an appropriate model for novelty detection, in: Fifth International Conference on Artificial Neural Networks (Conf. Publ. No. 440), 1997, pp. 117–122.

[26] L. Tarassenko, P. Hayton, N. Cerneaz, M. Brady, Novelty detection for the identification of masses in mammograms, in: 1995 Fourth International Conference on Artificial Neural Networks, 1995, pp. 442–447.

[27] D.-Y. Yeung, C. Chow, Parzen-window network intrusion detectors, in: Object recognition supported by user interaction for service robots, Vol. 4, 2002, pp. 385–388 vol.4.

[28] I. Ali, G. Saha, A distance metric based outliers detection for robust automatic speaker recognition applications, in: 2011 Annual IEEE India Conference, 2011, pp. 1–4.

[29] X. Dai, M. Bikdash, Distance-based outliers method for detecting disease outbreaks using social media, in: SoutheastCon 2016, 2016, pp. 1–8.

[30] K. Z. Haigh, L. M. Kiff, G. Ho, The independent lifestyle assistant: Lessons learned, Assistive Technology 18 (1) (2006) 87–106.

[31] S. S. Khan, M. G. Madden, One-class classification: taxonomy of study and review of techniques, The Knowledge Engineering Review 29 (3) (2014) 345–374.

[32] F. T. Liu, K. M. Ting, Z. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413–422.

[33] J. R. Pasillas-Díaz, S. Ratté, An unsupervised approach for combining scores of outlier detection techniques, based on similarity measures, Electronic Notes in Theoretical Computer Science 329 (2016) 61 – 77.

[34] G. Dib, O. Karpenko, E. Koricho, A. Khomenko, M. Haq, L. Udpa, Ensembles of novelty detection classifiers for structural health monitoring using guided waves, Smart Materials and Structures 27 (1) (2018) 15003.

[35] M. Mahmud, M. S. Kaiser, M. M. Rahman, M. A. Rahman, A. Shabut, S. Al-Mamun, A. Hussain, A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications, Cognitive Computation 10 (5) (2018) 864–873.

[36] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, Springer Publishing Company, Incorporated, 2014.

[37] D. J. Cook, N. Krishnan, Mining the home environment, Journal of Intelligent Information Systems 43 (3) (2014) 503–519.

[38] D. J. Cook, A. S. Crandall, B. L. Thomas, N. C. Krishnan, Casas: A smart home in a box, Computer 46 (7) (2013) 62–69.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.