Developing Ensemble Methods for Detecting Anomalies in Water Level Data

Thakolpat Khampuengson[1,2], Anthony Bagnall[1], and Wenjia Wang[1]

[1] School of Computing Sciences, University of East Anglia, Norwich, United Kingdom
{T.Khampuengson, Anthony.Bagnall, Wenjia.Wang}@uea.ac.uk
[2] Hydro-Informatics Institute of Ministry of Higher Education, Science, Research and Innovation, Thailand
thakolpat@hii.or.th

Abstract. Telemetry is an automatic system for monitoring environments in a remote or inaccessible area and transmitting data via various media. Data from telemetry stations can be used to produce early warning or decision supports in risky situations. However, sometimes a device in a telemetry system may not work properly and generates some errors in the data, which can then cause false alarms or miss true alarms for disasters. To detect such errors, human experts are usually required to investigate the data but this manual process is not only experience-dependent but also very time-consuming. In this paper, we present two ensemble methods for automatically detecting the anomaly in telemetry water level data. The novelty of our methods is that, although the individual models are conventional, by combining some of these models selected with a combined scoring function, our ensembles are able to improve the anomaly detection accuracy and reliability. Two types of ensembles were developed - simple and complex ensembles. A simple ensemble is built with the models selected from 7 basic anomaly detection models, and a complex ensemble is built with the selected simple ensemble models. The ensembles were tested on the data collected from 17 water level stations and the results clearly show that the complex ensembles are most accurate and also reliable in detecting anomalies.

Keywords: Ensemble Methods · Water Level Telemetry Monitoring · Anomaly Detection.

1 Introduction

The accuracy of meteorological and hydrological information is essential for research, forecasting, decision making and environment protection. But such data can be corrupted with some errors during the process of collection and transmission, because in certain environments and circumstances, the data have to be collected with the instruments installed at remote stations and then transmitted to a data centre through a telemetry system.

The Hydro Informatics Institute (HII) [3], Thailand, is responsible for installing telemetry stations in Thailand, for monitoring water levels and developing flood early warning systems. A telemetry station is equipped with various instruments to collect hydrological and meteorological information such as temperature, humidity, air pressure, rainfall, and water level, etc. and then transmit the collected data to the HII data centre every 10 minutes through cellular or satellite networks. The data can be accessed online via a website [4].

The telemetry data can be analysed to produce early warnings and decision supports to the relevant government agencies for dealing with critical and risky situations, such as heavy rainfall or fast raising water level, which may cause flooding. However, sometimes the devices in the telemetry system went wrong and generated various errors in data. For example, there were cases where the data from a telemetry station reported that there was a heavy rainfall in the area, but when the data was verified, there was actually no rain in that area at that particular time, this event is called “false alarm”. Although, we can verify the data before dissemination, the process of detecting errors in the data requires experienced humans to investigate the data and make decisions. This human intervention process is time consuming given the huge quantity of the collected data, and also produces inconsistent decisions due to variations of human’s experience. All these issues can cause considerable delay in detecting abnormal water levels at the right time and location, and issuing an early warning for a flood situation that may occur in real time.

In this research, we aim to address these issues by developing some intelligent methods to detect anomalies in water level data automatically as fast and accurate as possible.

2 Related Works

The data generated from a water level monitoring station is of time series. There are several types of existing models to detect anomaly in time series data.

The choice of models depends on the type of anomaly. For example, for missing and outlier values, K-mean clustering method[15] was usually used as it is simple and relatively effective. Simple and exponential smoothing techniques were used to identify an anomaly in a continuous data stream of temperature in an industrial steam turbine[14]. Two-sided median method and the One-sided median method were used to detected unexpected jump values in time series[1].

Most of their models have been computed based on a sliding window [9, 6, 13]. A time window is specified with n continuous input data points, and a model has been generated from the data in this window. Then the window is shifted by a given step size along time series and the model is recomputed on the next window. This has two drawbacks: the computed values limited to a specific window and time-consuming.

[3] http://www.hii.or.th

[4] http://www.thaiwater.net

Some machine learning methods were used to detect anomalies. But as almost all the applied methods use supervised learning algorithms to learn from the labelled historical data, there are several issues with them. (a) There is not enough labelled data for a learning algorithm to learn well to generate good enough models. (b) As time-series time is continuous or streaming over in real time, the models learned from historical data may have to be retrained with new data arrive[12]. (c) Each model is limited by the data it has been trained with, so may be suitable to detect a particular type of errors, but not other anomalies. Some previous researches have shown that it is possible to combine various individually trained models to produce more accurate detections than any of the single models[17]. This combination of multiple models to work together is called ensemble method. It has been demonstrated to be effective in a widespread of real-life problems, such as weather forecasting[8], detecting anomalies in cellular networks[4], wireless sensor networks detection[5], gene expression data for cancer classification[18].

These successful studies motivated us to develop ensembles for our problem. Before describing our ensemble methods, we will introduce the data we used in this study.

3 Data and Types of Anomaly

As our objective is to develop algorithms for detecting anomalies in water level data from HII telemetry station, we chose the data collected from some water level stations in Thailand as our testing cases in this study.

3.1 Water Level Datasets

To demonstrate and compare the efficiency of each anomaly detection model that will be developed in this research, we chose 17 telemetry stations from Yom Basin that installed in the same year, which represents the typical geological, meteorological and hydrological features in Thailand. All the stations installed VEGA PULS WL 61/62 instruments to measure water level every 10 minutes during the years of 2013 to 2018.

The details of the data are summarised in Table 1. We can see that not every station has anomaly data and some stations only have one anomaly. Because of HII telemetry stations installed radar water level gauge located at a fixed place above the water surface, the object under the sensors such as weeds, boats, or things that float along the water can affect sensors' reading of water level. In a dry season, the sensor may not detect the water surface properly, instead, it may detect waterbed or grass land. Figure 1 gives an example of such, where the water level sensor gave abnormal or incorrect readings on the water levels in Summer but produced reasonable water levels for Rainy. For these reasons, the data from these stations are very difficult for human to label normal or abnormal as ground-truth. Although we can avoid the stations that have a low number of anomaly from this experiment, we want to keep it to test our models to see if

Table 1. Summary of data from 17 telemetry stations.

Station	No. of Records	No. of Anomaly	% of Anomaly Data
DIV002	167708	24106	14.37
DIV004	215330	121	0.06
DIV006	216541	20849	9.63
NAN008	113833	1	0.00
VLGE13	191276	57	0.03
YOM001	177323	0	0.00
YOM002	179756	0	0.00
YOM003	192959	0	0.00
YOM004	163131	92	0.06
YOM005	168782	2902	1.72
YOM006	149321	4	0.00
YOM007	170142	0	0.00
YOM008	162099	2	0.00
YOM009	178241	2626	1.47
YOM010	202398	3080	1.52
YOM011	134031	638	0.48
YOM012	224404	1	0.00

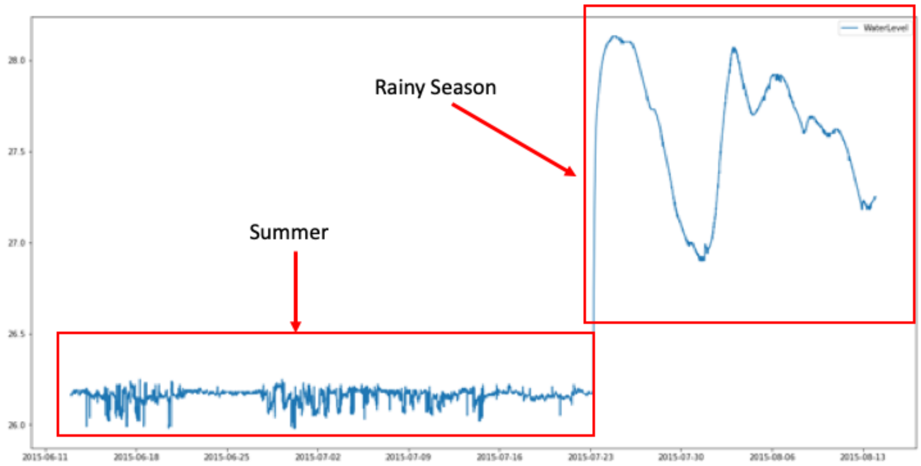

Fig. 1. An example of water level instrument that has been affected by the environments. The instrument gave unreasonable readings as water levels should not be oscillating abruptly by physics.

they can misidentify normal data, which in turn can help our experts to check the labelled anomaly data for validating the ground-truth.

3.2 Type of Anomaly Data

During the data collection and transmission, some errors can be introduced to the data. We analysed the data and classified the anomaly values into 4 types included. (1) *Missing value* is that not received or not transmitted from a station for the various reason. (2) *Spike* is the value that has a large and sharp difference from previous and following values in a sudden and short period of time. (3) *Pattern error* is an error value that repeats in a short or long-time period. (4) *Inconsistency error* is a discrepancy between the measured values when compared with a nearby station or other related values.

In this paper, we focus on detecting *spikes* because they occurred most frequently in our data and seriously affected the accuracy of early warning, so they need to be detected as soon as possible, and as accurate as possible.

3.3 Data Labelling

Water level data from telemetry stations is unlabelled and needs to be assigned with the ground-truth for model's learning and evaluation. The experts at the HII were used to analyse the data and to identify various anomalies in our hydrological data. Their decisions were aggregated to produce a consensus label for each possible anomaly.

4 Ensemble Methods

From the literature review, we have identified 7 classic methods for detecting anomalies. In general they are simple and fast so have been used in many applications, but each model has its limits, only suitable for detecting some particular types of errors. Nevertheless, they can be constructively combined into form an ensemble to work together so that they can compensate each other's weakness and then perform better than individual models working separately. One of the co-authors has studied the fundamental issues of ensemble methods [19] and emphasised that a successful ensemble can be built with some appropriate models selected by using suitable criteria, otherwise an ensemble may not improve at all. So, in this research, we developed two types of ensembles. The first type is a *simple ensemble*, the second type is the *complex ensemble* which has a compound structure of ensemble of ensembles. This section describes how these ensembles are constructed in detail.

Before introducing our ensemble methods, we briefly describe the conventional models that we chose as the candidates to be selected to build an ensemble.

4.1 Basic Anomaly Detection Model

Seven basic models were selected as the member candidates for building an ensemble, which are *Auto Regression(AR), Differenced Based(DB), Interquartile Range(IQR), Sigma rules of thumb(K-Sigma), Moving Average Smoothing(MAS), Slope as an Angle(SA), Z-Score.* They were chosen because they are simple and use no or few data for training. As a result, they take much shorter time to calculate, thus are suitable for detecting anomaly in near real-time situations. Each of these 7 models is briefly summarised below.

- *Autoregression(AR)*
  AR is an autoregressive modelling method that can predict future values from the previous values in a time series [11]. Therefore, the new data that more or lower than the accuracy interval of prediction should be abnormal.
- *Difference Based (DB)*
  In time series data, the difference between a normal and an abnormal value is usually more than the average difference of previously normal values. Therefore we can use the difference between the two observations to find the anomaly [3].
- *Interquartile Range (IQR)*
  An outlier or extreme value can be detected by using median as opposed to the mean, which is often summarized by the difference between the first and the third quartiles, well-known as the *Interquartile Range(IQR)* [16], the values that are not in this range will be defined as an anomaly.
- *Sigma Rule of Thumb (K-Sigma)*
  A data set that has most of the data distributed around the mean values in a symmetric shape is called *"normally distributed"*. In principle, when a data set has a normal distribution then the mean and the median are the same (in case of perfectly bell-shape) or almost equal, 99.7% of all data fall between $[mean \pm 3 * \sigma]$ the last rule is also know as "three-sigma rule of thumb" [10]. We defined values outside that interval to be anomalous.
- *Moving Average Smoothing (MAS)*
  For a given time point t in a time series data, this technique firstly defines a window around it, and then calculates the average of the raw observations in the window and uses the average as the threshold of that time point. User has to define the number of raw observation data, called windows size. The anomaly has been identified by comparing the expected values and a threshold is calculated. This method is used to remove the noise from the data. It is simple and commonly used in time series analysis and forecasting [2].
- *Slope as Angle (SA)*
  The anomaly data is the point that is usually significantly different from the previous data point. Consequently, the value that could be anomaly will have a slope angle close to 90°. In this research, we defined a point as an anomaly if there is angle slope more than 45°.
- *Z-Score*
  Z-score is the difference between the value and the mean expressed as the number of standard deviations. A observation value that has a Z-score lower than -2.5 and greater than 2.5, it will be considered as anomaly [7].

4.2 A Modified Sliding Window Algorithm

Every model, except SA, requires to employ a *sliding window* algorithm to find the threshold $\theta(w_t)$ for the current window w_t and uses it to decide if the value $x(t)$ is an anomaly or not. But it has a drawback, that is, when the window is moved forward to the next step, the detected anomaly data will be included in the window and because this anomaly data has not been removed or corrected, it will add some bias to the threshold and then as a consequence it will affect the prediction of next possible anomaly value along the time series.

We then modified this basic sliding window algorithm to address these issues. The basic idea is to remove the identified anomaly $x(t)$ and replace it with the value $x(t+1)$ so that the anomaly will not affect the threshold value of the next window, w_{t+1}.

On the other hand, if $x(t)$ has been verified as normal, then the window moves forwards as normal. We named our modified algorithm as the *Only Normal Sliding Windows(ONSW)* and the algorithm is given in algorithm 1.

Algorithm 1 Only Normal Sliding Windows

```
Set $n$ and $t$ equal to the size of window
$w_t = \{x(t-n), x(t-n+1), ..., x(t-1)\}$                 ▷ Set initial values for $w_t$
while $t$ less than count of data in $x$ do
    $\theta(w_t) = Model(w_t)$                  ▷ Finding the threshold from selected model
    if $x(t) > \theta(w_t)$ then                              ▷ Compare with threshold
        $x(t)$ is anomaly                                     ▷ Detect $x(t)$ is anomaly
    else
        $w_t = \{x(t-n+1), x(t-n+2), ..., x(t)\}$             ▷ Update data in window
    $x(t) = x(t+1)$                                            ▷ Move to next data
```

4.3 Criteria for selecting models

In order to select some models to build an ensemble, some measures should be chosen as criteria to evaluate the accuracy of models. As this is essentially a binary classification problem, we chose the values from the confusion matrix which is a widely used technique for summarizing the performance of a classification algorithm. As our purposes are to detect the anomaly and reduce the false alarm, TP, FP and FN are suitable to use as criteria. Where TP, FP and FN denote the number of True Positive - correct predictions for anomaly data, False Positive - the number of incorrect predictions for anomaly data, False Negative - the number of incorrect predictions for normal data, respectively.

4.4 Simple ensemble

A simple ensemble is built with some models selected from 7 basic models briefly mentioned earlier. But a key question is what criterion we should use to select a model as a member of an ensemble. We devised a new scoring function (see below) to calculate the goodness score of a model and then use this score to determine if a model is good enough to be selected. Once an ensemble is formed, a decision-making function is applied to work out the final output of the ensemble. In this study, a simple majority voting method is used. So, a simple ensemble operates in 2 main stages: *Model Selection* and *Decision Making.*

- *Model Selection*: It is done in three steps:
  (1) Evaluating the accuracies of models.
  We firstly use three different measures TP, FP and FN as criteria to asses the performance of a model.
  (2) Ranking the models with different criteria.
  With those 3 measures, we produce 3 rankings R_1, R_2, and R_3 respectively. Then for each ranking of the models, we calculate their score in ranking R_j as follows:

$$S_{(m_i,R_j)} = \frac{N+1-r_{(m_i,R_j)}}{N}, \in [\frac{1}{N}, 1] \quad (1)$$

  Where,

  $S_{(m_i,R_j)}$ = Score of model m_i in ranking R_j.
  $r_{(m_i,R_j)}$ = ranking position of m_i in R_j.
  N = number of models in a ranking.
  i = index of models: 1, 2, ..., N.
  j = index of Rankings: 1, 2, 3.

  Then we devised a new measure - Total Score of Performance(TSP), that combines the three scores from each model (TSP_{m_i}) by the following equation.

$$TSP_{m_i} = 1/N \sum_{j=1}^{3} S_{(m_i,R_j)}, \in [\frac{1}{N}, 1] \quad (2)$$

  Then all the models are ranked again by their TSP score in a descending order, i.e. the models with higher TSP values are ranked higher. In doing so, we produced 4 rankings for each model.
  (3) Selecting models for building ensembles
  In this stage, we need to decide what models and how many models should be selected to build an ensemble. A general consideration is to select a certain number of suitable models that will maximize the performance of an ensemble model. To avoid a tie-situation in decision making, we set the number of member models to be an odd number as there are only 7 basic models in total in this experiment, we set three different sizes for ensembles: 7, 5 and 3, to investigate whether the size of an ensemble has any influence on its performance. So we chose top 7, 5 and 3 models from each ranking to build simple ensembles respectively. In this way, we built 9 simple ensembles and they are coded with their size and the used measure, e.g. Top5TP represents an ensemble built with top 5 of TP score models. In summary, we have 4 ensembles with top 3 models from each rankings, i.e.Top3TP, Top3FN, Top3FP and Top3TSP, 4 with top 5 models from each rankings: Top5TP, Top5FN, Top5FP and Top5TSP, and one ensemble with all the 7 models, called Ensem7.
  In addition, we also used another pair of measures - $Sensitivity$ and $Specificity$ to select the same numbers of models to build ensembles for comparison.

So, we have 4 more simple ensembles, Top5Sen, Top3Sen, Top5Spec and Top3Spec.

In total, we constructed 13 simple ensembles based on 6 different performance measures and 2 different sizes.

- *Decision Making*: Although there are several decision-making strategies to combine the outputs of the models in an ensemble, in this research we chose the simple majority voting approach for its simplicity and efficiency, which is particularly essential for our anomaly detection system to work fast enough in real-time with streaming data. A data point will be classified as an anomaly by an ensemble if more than half of its member models predict it as anomaly, otherwise, normal.

These simple ensembles were tested on the testing data and the results will be presented in the next section. Our initial experimental results show that simple ensembles are better than the individual models but we want to improve further. So we developed a method for building complex ensembles.

4.5 Complex ensemble

A complex ensemble is built by using selected simple ensembles as its member models. It can simply be viewed as an ensemble of ensembles, so donated as *EoE*. An EoE still uses the majority voting among the selected simple ensembles to determine its final result. From the 13 simple ensembles built earlier, we can construct 5 complex ensembles by selecting top 3, 5, 7, 9 and 11 simple ensembles based on their TSP score, and another one with all the 13 simple ensembles. They are donated as EoE3, EoE5, EoE7, EoE9, EoE11 and EoE13, respectively.

These complex ensembles were in the same ways as for the individual models, and simple ensembles with the same dataset. The results are described in the next section.

5 Experiments and Results

The 7 classic anomaly detection methods have been trained with the training data of the chosen stations by moving the windows over the entire duration to find their decision threshold. Then those cut-off values are applied to the testing data to evaluate their accuracy with $Recall$, $Precision$, and $F1$ scores, by comparing predicted values and their corresponding ground-truth. Then these basic individual models were used to build simple ensembles and the simple ensembles were used to build complex ensembles. All the ensembles were tested in the same manners as for the individual models. Their testing results are presented below separately.

5.1 Individual Model Results

The total scoring performance (TSP) on 17 stations from each model is presented in Table 2. Although the SA model has the highest TSP score at 1.00, further examinations found that it occurs in the datasets that have few or none of anomaly data. The AR model has the lowest performance with an average TSP score of only 0.46. The IQR model is the best model because it has not only the highest score from half of all the 17 stations but also the highest total and average scores over all the stations.

We also calculated $Recall, Precision$, and $F1$ scores for each model and the results of the best model IQR in terms of $Recall$ score are given in Table 3. It should be noted for 4 stations YOM001, YOM002, YOM003 and YOM007 because anomaly data have not been labelled in their datasets, it is not meaningful to compute these three measures.

The results show that the even best individual model performed quite badly in terms of $Precision$ and $F1$, only got the overall averages of 20.56% and 34.39% respectively. Moreover, it can be seen that it performed very poor on some stations including DIV004, NAN008, YOM006, YOM008, and YOM012.

Table 2. Total and average weighting score of individual model

Station	AR	DB	IQR	KSigma	MAS	SA	ZScore
DIV002	0.43	0.43	**0.86**	0.43	0.76	0.43	0.76
DIV004	0.24	0.62	**0.81**	0.57	0.67	0.43	0.76
DIV006	0.52	0.52	**0.76**	0.33	0.67	0.43	0.76
NAN008	0.33	0.57	0.86	0.81	**0.90**	0.62	0.76
VLGE13	0.43	0.48	**0.90**	0.48	0.81	0.52	0.76
YOM001	0.71	0.95	0.86	0.90	0.76	**1.00**	0.81
YOM002	0.71	0.90	0.86	**1.00**	0.76	0.95	0.81
YOM003	0.71	0.95	0.86	0.90	0.76	**1.00**	0.81
YOM004	0.43	0.43	0.86	0.43	**1.00**	0.48	0.76
YOM005	0.52	0.48	**0.90**	0.48	0.52	0.52	0.76
YOM006	0.14	0.95	0.81	0.38	0.90	**1.00**	0.76
YOM007	0.71	0.95	0.86	0.90	0.81	**1.00**	0.76
YOM008	0.33	0.57	**0.90**	0.81	0.86	0.62	0.76
YOM009	0.52	0.43	0.71	0.48	0.67	0.43	**0.76**
YOM010	0.33	0.62	**0.86**	0.48	0.62	0.43	0.67
YOM011	0.43	0.48	**0.81**	0.48	0.71	0.52	0.76
YOM012	0.33	0.57	0.81	0.86	**0.90**	0.62	0.76
Total	7.86	10.90	**14.29**	10.71	13.10	11.00	13.00
Average	0.46	0.64	**0.84**	0.63	0.77	0.65	0.76
Std.	0.17	0.21	0.05	0.23	0.12	0.24	0.03

5.2 Simple Ensemble Results

The testing results of 13 simple ensembles we built are given in Table 4. They show that the simple ensembles are generally better than individual models. Specifically, Top5TP, Top5FN, Top5Sen, and Top3TSP have the highest total scores (12.95) and the average score of 0.76. But when we looked at them in more detail, we found that Top5TP, Top5FN and Top5Sen have never been the best models in any station, and in contrast, Top3TSP has the highest performance in 6 stations. Although Top3FP and Top3Spec have the highest score, 1.00, in 5 stations, these stations have actually no anomaly.

Table 3. The results from IQR Model for each station.

Station	IQR						
	TP	FN	FP	TN	Recall	Precision	F1
DIV002	24105	0	907	142696	1.0000	0.9637	0.9815
DIV004	121	0	10415	204794	1.0000	0.0115	0.0227
DIV006	18989	1860	4970	190722	0.9108	0.7926	0.8476
NAN008	1	0	413	113419	1.0000	0.0024	0.0048
VLGE13	57	0	2191	189028	1.0000	0.0254	0.0495
YOM001	0	0	2297	175026	-	-	-
YOM002	0	0	2440	177316	-	-	-
YOM003	0	0	1737	191222	-	-	-
YOM004	92	0	610	162429	1.0000	0.1311	0.2317
YOM005	2902	0	2540	164350	1.0000	0.5333	0.6956
YOM006	4	0	1707	147610	1.0000	0.0023	0.0047
YOM007	0	0	394	169748	-	-	-
YOM008	2	0	228	161869	1.0000	0.0087	0.0172
YOM009	2624	2	9673	165942	0.9992	0.2134	0.3517
YOM010	2983	101	10017	189297	0.9673	0.2295	0.3709
YOM011	615	0	2122	131294	1.0000	0.2247	0.3669
YOM012	1	0	1730	222673	1.0000	0.0006	0.0012
Average					0.9912	0.2056	0.3439
Std.					0.0247	0.3359	0.3634

Table 4. Total and average TSP scores of simple ensemble models.

Station	Ensem7	Top5TP	Top3TP	Top5FN	Top3FN	Top5FP	Top3FP	Top5Sen	Top3Sen	Top5Spec	Top3Spec	Top5TSP	Top3TSP
DIV002	0.49	0.59	**0.77**	0.59	**0.77**	0.59	0.44	0.59	**0.77**	0.59	0.44	0.67	**0.77**
DIV004	0.49	0.67	**0.77**	0.67	**0.77**	0.54	0.44	0.67	**0.77**	0.54	0.44	0.59	**0.77**
DIV006	0.54	0.67	0.69	0.67	0.69	0.54	0.44	0.67	0.69	0.54	0.44	0.28	**0.77**
NAN008	**0.95**	0.87	0.77	0.87	0.77	**0.95**	0.44	0.87	0.77	**0.95**	0.44	0.87	0.69
VLGE13	**0.90**	0.85	0.77	0.85	0.77	0.49	0.44	0.85	0.77	0.49	0.44	0.87	0.77
YOM001	0.95	0.82	0.74	0.82	0.74	0.95	**1.00**	0.82	0.74	0.95	**1.00**	0.87	0.85
YOM002	0.90	0.87	0.74	0.87	0.74	0.95	**1.00**	0.87	0.74	0.95	**1.00**	0.79	0.77
YOM003	0.90	0.82	0.74	0.82	0.74	0.95	**1.00**	0.82	0.74	0.95	**1.00**	0.87	0.85
YOM004	0.59	0.85	0.77	0.85	0.77	0.59	0.59	0.85	0.77	0.59	0.59	**0.87**	0.69
YOM005	0.49	0.62	0.74	0.62	0.74	0.49	0.44	0.62	0.74	0.49	0.44	0.51	**0.85**
YOM006	**1.00**	0.87	0.77	0.87	0.77	0.92	**1.00**	0.87	0.77	0.92	**1.00**	0.79	0.69
YOM007	**1.00**	0.85	0.74	0.85	0.74	**1.00**	**1.00**	0.85	0.74	**1.00**	**1.00**	0.87	0.77
YOM008	0.90	0.82	0.74	0.82	0.74	**1.00**	0.44	0.82	0.74	**1.00**	0.44	0.85	0.87
YOM009	0.49	0.64	0.74	0.64	0.74	0.38	0.54	0.64	0.74	0.38	0.54	0.51	**0.77**
YOM010	0.49	0.64	**0.74**	0.64	**0.74**	0.49	0.44	0.64	**0.74**	0.49	0.44	0.51	0.62
YOM011	0.49	0.64	0.74	0.64	0.74	0.49	0.44	0.64	0.74	0.49	0.44	0.67	**0.77**
YOM012	0.90	0.87	0.77	0.87	0.77	**0.95**	0.44	0.87	0.77	**0.95**	0.44	0.87	0.69
Total	12.44	**12.95**	12.77	**12.95**	12.77	12.26	10.49	**12.95**	12.77	12.26	10.49	12.28	**12.95**
Average	0.73	**0.76**	0.75	**0.76**	0.75	0.72	0.62	**0.76**	0.75	0.72	0.62	0.72	**0.76**
Std.	0.22	0.11	0.02	0.11	0.02	0.24	0.26	0.11	0.02	0.24	0.26	0.18	0.07

As can be seen, Top3TSP is the best simple ensemble model. Moreover, when we considered binary performance values as presented in Table 5, we observed that when the TP score increases, the FN and FP values decrease. Especially when the FP reduces more than 50% in some stations (e.g. DIV002, DIV006, and YOM010), but the average *Recall* still remains at 99% whilst the average *Precision* has been increased by 8% when compared with the best individual model, the IQR.

Table 5. Classification accuracies of Top3TSP simple ensemble.

Station	Top3TSP						
	TP	FN	FP	TN	Recall	Precision	F1
DIV002	24105	0	256	143347	1.0000	0.9895	0.9947
DIV004	121	0	10588	204621	1.0000	0.0113	0.0223
DIV006	19026	1823	2710	192982	0.9126	0.8753	0.8936
NAN008	1	0	306	113526	1.0000	0.0033	0.0065
VLGE13	57	0	1268	189951	1.0000	0.0430	0.0825
YOM001	0	0	132	177191	-	-	-
YOM002	0	0	328	179428	-	-	-
YOM003	0	0	85	192874	-	-	-
YOM004	92	0	64	162975	1.0000	0.5897	0.7419
YOM005	2902	0	1336	165554	1.0000	0.6848	0.8129
YOM006	4	0	290	149027	1.0000	0.0136	0.0268
YOM007	0	0	48	170094	-	-	-
YOM008	2	0	40	162057	1.0000	0.0476	0.0909
YOM009	2624	2	2800	172815	0.9992	0.4838	0.6519
YOM010	2931	153	4148	195166	0.9504	0.4140	0.5768
YOM011	615	0	1265	132151	1.0000	0.3271	0.4930
YOM012	1	0	467	223936	1.0000	0.0021	0.0043
Average					0.9902	0.2815	0.4517
Std.					0.0259	0.3700	0.3939

Table 6. Total and average TSP scores of complex ensembles.

Station	EoE13	EoE11	EoE9	EoE7	EoE5	EoE3
DIV002	0.56	0.56	0.56	**0.83**	**0.83**	**0.83**
DIV004	**0.89**	**0.89**	0.67	0.72	**0.89**	**0.89**
DIV006	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
NAN008	0.89	0.89	0.89	0.89	**1.00**	**1.00**
VLGE13	0.94	0.94	0.72	0.94	0.94	**1.00**
YOM001	0.94	0.94	0.94	0.94	0.94	**1.00**
YOM002	0.83	0.83	0.83	0.94	0.94	**1.00**
YOM003	0.83	0.83	0.83	0.94	0.94	**1.00**
YOM004	0.72	**1.00**	0.44	**1.00**	0.83	0.83
YOM005	0.56	0.50	0.56	**0.83**	0.78	0.78
YOM006	0.78	0.78	**1.00**	**1.00**	**1.00**	**1.00**
YOM007	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
YOM008	0.83	0.83	0.83	0.89	0.94	**1.00**
YOM009	0.61	**0.78**	0.67	0.72	**0.78**	**0.78**
YOM010	0.44	0.50	0.56	**0.83**	0.78	0.78
YOM011	0.67	0.61	0.78	0.78	0.61	**0.83**
YOM012	0.89	0.89	0.89	0.89	0.94	**1.00**
Total	13.39	13.78	13.17	15.17	15.17	**15.72**
Average	0.79	0.81	0.77	0.89	0.89	**0.92**
Std.	0.17	0.17	0.18	0.09	0.11	0.10

5.3 Complex Ensemble Results

Table 6 shows the scores of complex ensembles. It is clear that EoE3 is the best with an overall average TSP score of 0.92. It is followed by EoE5 and EoE7 with the same average TSP scores at 0.89. In addition, EoE3 has the best performance in 14 stations, out of 17, and produced the full score for 10 of these 14 stations.

Table 7 gives the detailed measures for EoE3. It shows that although the average *Recall* score of EoE3 decreased a bit, it has achieved the highest average *Precision* score, which is 5% more than the best simple ensemble Top3TSP and 13% more than the best individual model IQR. Especially, the number of False Positive predictions has reduced significantly.

It is worth mentioning that for stations DIV006, YOM009, and YOM010, all the models predicted high FN values. We took a close look at them and found

Table 7. Anomaly detection results from complex ensemble EoE3.

Station	EoE3						
	TP	FN	FP	TN	Recall	Precision	F1
DIV002	24105	0	256	143347	1.0000	0.9895	0.9947
DIV004	120	1	1434	213775	0.9917	0.0772	0.1433
DIV006	18482	2367	728	194964	0.8865	0.9621	0.9227
NAN008	1	0	15	113817	1.0000	0.0625	0.1176
VLGE13	57	0	55	191164	1.0000	0.5089	0.6746
YOM001	0	0	2	177321	-	-	-
YOM002	0	0	2	179754	-	-	-
YOM003	0	0	1	192958	-	-	-
YOM004	92	0	26	163013	1.0000	0.7797	0.8762
YOM005	2902	0	21521	145369	1.0000	0.1188	0.2124
YOM006	4	0	11	149306	1.0000	0.2667	0.4211
YOM007	0	0	6	170136	-	-	-
YOM008	2	0	11	162086	1.0000	0.1538	0.2667
YOM009	1984	642	2073	173542	0.7555	0.4890	0.5937
YOM010	3003	81	5555	193759	0.9737	0.3509	0.5159
YOM011	614	1	418	132998	0.9984	0.5950	0.7456
YOM012	1	0	27	224376	1.0000	0.0357	0.0690
Average					0.9715	0.3312	0.5358
Std.					0.0691	0.3570	0.3339

that the data from these stations have high fluctuations and as shown in Figure 1 on the Summer periods. So their data varied considerably from the real situations and thus are very difficult for human to determine their ground-truth. Then as a consequence, the models might not be able to learn and generalise well on the data from these stations. But some models performed well in identifying normal data, which is useful to keep the collected valuable data in the datasets. In addition, the normal data in some stations i.e. DIV004, NAN008, and YOM012, are greatly outnumbered by the anomalies, which leads to the low *Precision* values and high standard deviations.

6 Conclusions

In this research, we developed 2 types of ensemble models, *Simple* and *Complex* ensemble models, with 7 basic conventional models, for detecting anomaly in water level data from telemetry systems. We produced a modified a sliding window algorithm and devised a total scoring function(TSP) by combining 3 measures - TP, FP and FN, to assess the overall performance of a model. A simple ensemble is built with the models selected from the 7 classic models with a variety of selection criteria. A complex ensemble is built with the selected simple ensembles.

The *classic models, simple ensembles* and *complex ensembles* were tested on the data from 17 stations, the results show that the classic model IQR is the best individual model at detecting anomalies but poor for classifying normal data. In general simple ensembles are more accurate and consistent than individual models. The best simple ensemble, Top3TSP, outperformed the best individual model IQR by achieving the same accuracy on detecting anomaly data and more accurate results for normal data than IQR. Further improvements were produced by our complex ensembles. It is clear that the complex ensemble EoE3, with only three member models, beats both the best individual model IQR and the best simple ensemble Top3TSP with clear margins in detecting anomalies and also normal data. This is confirmed with the highest $F1$ score.

In conclusion, the developed ensemble methods can select some suitable basic individual models to build simple and complex ensembles to improve the accuracy of detecting anomalies in water level data. Our testing results demonstrated that our ensemble methods have a real potential to be further developed to help the related organisation HII to reduce their time in investigating the data and to improve the performance of early warning systems and decision support system. They can also be used to develop the firmware of telemetry station to be able to detect anomaly values by itself. In addition, we can apply the models to assist experts in labelling the data as ground-truth by comparing the results from our models.

The next step of our research is to further develop ensemble methods by integrating some appropriate machine learning methods, such as deep reinforcement learning for detecting other types of anomalies and then correcting anomalous data.

Acknowledgement

The authors would like to thank the Hydro-Informatics Institute of Ministry of Higher Education, Science, Research and Innovation, Thailand, for providing the scholarship for Mr. Thakolpat Khampuengson to do his PhD at the university of East Anglia and also his colleagues at the HII for labelling the data used in this research.

References

1. Basu, S., Meckesheimer, M.: Automatic outlier detection for time series: an application to sensor data. Knowledge and Information Systems **11**(2), 137–154 (2007)
2. Bernacki, J., Kołaczek, G.: Anomaly detection in network traffic using selected methods of time series analysis. IJ Computer Network and Information Security **9**, 10–18 (2015)
3. Chen, X.y., Zhan, Y.y.: Multi-scale anomaly detection algorithm based on infrequent pattern of time series. Journal of Computational and Applied Mathematics **214**(1), 227–237 (2008)
4. Ciocarlie, G.F., Lindqvist, U., Nováczki, S., Sanneck, H.: Detecting anomalies in cellular networks using an ensemble method. In: Proceedings of the 9th international conference on network and service management (CNSM 2013). pp. 171–174. IEEE (2013)
5. Curiac, D.I., Volosencu, C.: Ensemble based sensing anomaly detection in wireless sensor networks. Expert Systems with Applications **39**(10), 9087–9096 (2012)
6. Ding, Z., Fei, M.: An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proceedings Volumes **46**(20), 12–17 (2013)

7. Ghosh, D., Vogt, A.: Outliers: An evaluation of methodologies. In: Joint statistical meetings. vol. 2012 (2012)
8. Gneiting, T., Raftery, A.E.: Weather forecasting with ensemble methods. Science **310**(5746), 248–249 (2005)
9. Golab, L.: Querying sliding windows over online data streams. In: International Conference on Extending Database Technology. pp. 1–11. Springer (2004)
10. Grafarend, E., Awange, J.: Linear and Nonlinear Models: Fixed effects, random effects, and total least squares. Springer (2012)
11. Hill, D.J., Minsker, B.S.: Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environmental Modelling & Software **25**(9), 1014–1022 (2010)
12. Hill, D.J., Minsker, B.S., Amir, E.: Real-time bayesian anomaly detection for environmental sensor data. In: Proceedings of the Congress-International Association for Hydraulic Research. vol. 32, p. 503. Citeseer (2007)
13. Jiang, D., Liu, J., Xu, Z., Qin, W.: Network traffic anomaly detection based on sliding window. In: 2011 International Conference on Electrical and Control Engineering. pp. 4830–4833. IEEE (2011)
14. Kumar, A., Srivastava, A., Bansal, N., Goel, A.: Real time data anomaly detection in operating engines by statistical smoothing technique. In: Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on. pp. 1–5. IEEE (2012)
15. Lin, J., Sheng, G., Yan, Y., Zhang, Q., Jiang, X.: Online monitoring data cleaning of transformer considering time series correlation. In: 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D). pp. 1–9. IEEE (2018)
16. Mood, A.M.: Introduction to the theory of statistics. (1950)
17. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of artificial intelligence research **11**, 169–198 (1999)
18. Tan, A.C., Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification (2003)
19. Wang, W.: Some fundamental issues in ensemble methods. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 2243–2250. IEEE (2008)