Predicting failures in electronic water taps in rural sub-Saharan African communities: an LSTM-based approach

N. M. Offiong, Y. Wu and F. A. Memon

ABSTRACT

There is a growing need to sustain solar-powered water taps in most parts of the sub-Saharan Africa. The frequent failure of the water taps gives rise to intermittent water supply and poor service delivery by the water service providers. The challenge is to foresee and predict the failure of these water systems before they occur. This study develops a scalable machine-learning model for failure prediction in electronic water taps to ensure timely maintenance of the taps. Specifically, we develop a model based on long short-term memory (LSTM) to efficiently make failure predictions with noisy heterogeneous time-series data from rural water taps. Results from the experiment prove that the proposed model can effectively classify activities and patterns in various time-series datasets. With the proposed model, the failures of the solar-powered taps due to abnormal events can be successfully predicted well in advance, with an accuracy of 78.54%. Based on the data analyses, common causes of failures are presented.

Key words | anomaly detection, deep learning, failure prediction, LSTM, time-series data

N. M. Offiong (corresponding author)
F. A. Memon
The University of Exeter, Centre for Water Systems,
University of Exeter,
Exeter, EX4 4QF,
UK
E-mail: *no270@exeter.ac.uk*

Y. Wu
Department of Computer Science,
EMPS, University of Exeter,
Exeter, EX4 4QF,
UK

HIGHLIGHTS

- Detecting failures in solar-powered standalone taps using machine learning.
- Adaptive framework for data feature extraction.
- Failure prediction in rural water supply taps.
- Coding errors to check their frequency of occurrence.
- Sustaining rural water supply taps.

INTRODUCTION

Rural villages in most of sub-Saharan Africa depend on hand pumps and solar-powered water taps for clean domestic water supply (Foster & Cota 2014). These solar-powered taps, which are not adequate for the growing rural population, sometimes break down and typically cause a shortage in the water supply. The failure of these taps affects both water users and water service providers, and is a considerable concern for efficient ways to manage them for continuous service delivery. Hence, this research focuses on the need to develop a useful model for failure prediction in solar-powered water taps deployed in the sub-Saharan region.

This research employs a data-driven approach for failure prediction; this involves the use of the data generated from the case study to predict possible failures in solar-powered water taps. Some studies have used statistical analysis to investigate time series (i.e., data points with temporal ordering) based on the data acquired from an operational system (Cabrera 2007; Mazumder *et al.* 2019). Recently, the application of machine learning (ML) paradigms to manage

doi: 10.2166/wst.2020.542

water infrastructures has produced significant improvements in failure prediction by modelling temporal datasets acquired from the systems of study (Mounce *et al.* 2015). The advancements in ML paradigms have made it possible (Mounce *et al.* 2015), and the advancement in computing architecture has also made it possible for researchers to carry out failure prediction on time-series data (Jordan & Mitchell 2015).

Our failure prediction approach uses long short-term memory (LSTM), an ML-based approach, to analyse the time-series data obtained from stand-alone solar-powered electronic water taps. Besides the LSTM model, there are traditional methods for time series forecasting, which include the autoregressive (AR), autoregressive integrated moving average (ARIMA), K nearest neighbour (KNN) and the support vector Machines (SVM) models. These traditional methods can be categories into two groups, namely model-based and distance-based.

The model-based time-series classification approaches include the AR model, which is one of the simplest methods that have been used for time-series analyses by a lot of analysts (Kini & Sekhar 2013). However, the AR model has a disadvantage, which is that it requires the time-series to satisfy stationary assumptions. In practice, this requirement is always breached. Also, the ARIMA harnesses previously observed data (AR) and the moving average of past data points to model output indirectly by integrating differenced output values. Stationary time series are derived from this process, and are needed because of their ability to extrapolate information generated from a one-time step to the next (Sina & Thomas 2019). ARIMA also has some drawbacks, including the fact that it is limited when dealing with complex datasets (Chen & Wang 2019).

Time-series models based on distance include KNN and SVM. The two models can be directly applied to time-series classification. However, these models' optimization problem falls into non-convex optimization and can quickly encounter local minima. Optimizing these models can cause overfitting and may prove to be a difficult problem to solve (Fu *et al.* 2016). Another challenge with SVM is that it shows less efficiency for a complex and massive dataset, which is a characteristic of the dataset of this research.

However, the choice of LSTM hinges on recent advancements in computational power and its ability to learn from massive datasets with longer temporal sequences. Therefore, LSTM offers a better classification accuracy over these traditional methods when dealing with large amounts of time-series data (Nakisa *et al.* 2018). LSTMs also have a special memory built into the LSTM architecture, which allows it to store information for a more extended period (Shewalkar *et al.* 2019). The time-series data are generated through daily usage of the water taps; the data are collected and sent remotely to a base station where the water installations monitoring takes place. The data collected from the electronic water supply points (EWSP) are various time-series datasets that possess latent information about the system's behaviour and this information needs a systematic interpretation with the use of the LSTM approach (Laucelli & Giustolisi 2015).

Overall, the main target of this study is to identify technical failures in the EWSP by analysing the collected data and using the information obtained to predict EWSP failures. The failure prediction is made to support timely detection and system maintenance through the development of a data-driven warning system. We harnessed case studies from smart water setups in sub-Saharan Africa to validate the effectiveness of the developed system. Most of the recently conducted research focuses on the use of hydraulic and water-quality data to monitor failures (Wu & Rahman 2017; Mamun *et al.* 2020). In this study, we label anomalies in water withdrawal data and use the labels to determine which anomaly is likely to cause failures in a water dispensing system through the use of the LSTM model. This was achieved through preprocessing of historical water usage data, observing the possible causes of failure and using the LSTM model to predict their occurrence. Some of the potential failures considered in this paper include (i) failure due to lack of voltage in the system (indicating a failure in the solar energy supply system), (ii) malfunction of the electronic tap (due to mechanical failure), (iii) low flow error (due to shortage of water) and (iv) master tag error (due to chipset miscommunication).

The contribution of this study centres on the following:

- Anomaly prediction from solar-powered taps through the application of the LSTM paradigm. The proposed model is a classification framework based on LSTM and capable of adaptive data feature learning.
- The detection of different failure events by labelling anomalies in a stationary real-world dataset; in turn, the labelled anomalies are used to determine the type of failures that may occur and their frequency of occurrence.

Sustaining water withdrawal systems by predicting the water system's failures in rural areas of sub-Saharan Africa is one of the crucial issues that many researchers are diligently investigating (Behailu *et al.* 2017; Yuanyuan *et al.* 2017; Foster *et al.* 2018). In this section, we review already published theories and methods for failure prediction and

sustainability; and give a description of time series to guide the reader.

Time series, as described by Malhotra *et al.* (2015), is a set $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, where each of the points $x^{(1)} \epsilon R^m$ in the time series is an m-dimensional matrix $\{x_1^{(t)}, x_m^{(t)}, \ldots, x_m^{(t)}\}$ whose elements correspond to the input variables. Time series are datasets produced over time. The goal of forecasting procedures is to extract a meaningful amount of information from the time-series data during the training process (i.e., the time where a model learns trends in the data) and use the obtained information to predict the behaviour of the next time point with the aid of ML (Ram 2017). ML gives insight into the pattern created by the time series. In most cases, these are patterns that are not clear to humans because they are latent and too complicated for a standard statistical method to understand, hence the need to use the LSTM approach. (Muharemi *et al.* 2019).

Wu *et al.* (2015) used ML to manage smart water supply networks by simulating and optimizing the control of water distribution systems. The study showed that ML is capable of extracting salient data features from a large dataset through a layered architecture (Najafabadi *et al.* 2015; Chollet 2018). The study saw the development of a framework for ML to carry out predictive analysis and anomaly detection in time-series data with a specific focus on pressure, flow and water consumption.

In another study, three variants of ML techniques, including the LSTM method, were used to study a water system based on the time-series data generated by their case study (Wei *et al.* 2020). The study used the root mean square error (RMSE) and the coefficient of determination (R^2) to evaluate the model. The study showed that the LSTM model could provide a more precise and robust prediction than most of the other ML models by characterizing the lag in time between the external inputs and responses from the investigated studies.

Similarly, the research carried out by Xu *et al.* (2020) applied LSTM to predict failure related to pressure and some other abnormal operational conditions in a water distribution system. In that study, they considered three model inputs and these include the pressures at measuring points, the water supply pressure and the entry point flow of the water supply system. Liu *et al.* (2019) analysed and predicted water quality with the use of a LSTM network. In that study, they designed a model for drinking-water quality and used the model to make future predictions on the quality of water based on past datasets.

The related works reviewed from other researches concentrated more on flow, pressure and water quality. In our research, we are interested in labelling some of the anomalies associated with the electronic tap (e-tap) and make predictions based on the labels.

MATERIALS AND METHODS

Dataset

The datasets used in this study are historical data (January to December 2018) of real usage of the EWSP installed by a commercial water services provider in the Gambia. The data contains a record with 1,047,114 rows representing time-series samples from 27 different EWSPs and 22 columns describing the features of the dataset. The data were collected over one year and transmitted to a remote server where they are kept for processing and planning purposes. The data contains both relevant features and other features that are not relevant to this study. From the dataset, we extracted the relevant feature for the analysis and they are shown in Table 1.

The three most important data columns that helped to answer the overall aims of the project are *ErrorCode*, *Voltage* and *FlowRate*. We split the data into two parts (80% for training (part of which we used for validation) and 20% for testing. Following this technique, we trained the developed model (explained later) on all the available data by selecting the time series on a step-by-step basis. During the implementation, we designed the training set to expand after each iteration while the test set remains fixed at a one-time step. The training and validation of the model occurred inside the loop of each iteration. With each iteration, a new model is created. This procedure repeats for

Table 1 | Extracted relevant data feature for the time-series analysis

Column header	Column description
AssetID	Unique identification for the EWSPs (tap number)
ErrorCode	Different codes to indicate issues with water tap usage (e.g. F1, F2, F3, F4)
FlowCount	Derived value to indicate value for flow rate
FlowTime	Time at which flow occurred
Litres	Volume of water withdrawn
Voltage	The battery voltage capacity detected at each instance the water tap accessed
DateTime	Time of event (water withdrawal date and time)
FlowRate	Volume of water collected per unit time

all the sequences and the average taken to give the overall effectiveness of the model. The splitting of the data was done as follows:

1. Training set (for the network to learn from)
2. Testing set (to evaluate the model)
3. Validation set (for early stopping and optimization of the parameters that the proposed model cannot learn by itself)

The original dataset is a noisy dataset with some missing information and mismatched datatypes, and preprocessing was done on the data to normalize the dataset and prepare it for the analysis. The data normalization method used in this study was the min-max method, which transforms the minimum value of the data to 0 and the maximum value to 1. The values between the minimum and maximum values are transformed to decimal values between 0 and 1. The formula for min-max normalization is shown in Equation (1) below:

$$x_{(0,1)} = \frac{x - \min}{\max - \min} \tag{1}$$

where $x_{(0,1)}$ and x are the normalized and original datasets, respectively. Min and max are the minimum and maximum values of the whole dataset, respectively.

The min-max normalization algorithm is widely used (Patro & Sahu 2015). The drawback of the method includes inefficiency in handling outliers in datasets. To solve this problem, we expanded the min-max range to accommodate any outliers in the dataset. Before the proposed LSTM-based model was built, the dataset was transformed to match a supervised learning format needed to solve the time-series modelling problem, where the time steps were structurally transformed to input and output values.

An initial analysis was done on the data with a conventional data manipulation to show how the two columns, *Voltage* and *FlowRate*, relate to one another, and more importantly, how both columns affect the behaviour of the e-taps. The preliminary analysis in Figure 1 shows the failures and normal functioning of the investigated taps.

Figure 1 reveals some failures and normal functioning of taps, but our interest is in failure trends. It can be seen from the preliminary analysis that at a point where the voltage of the system is 0, there was a considerable amount of flow from the EWSPs. It can also be seen that at a point where the voltage was at a peak, there was a failure and the tap did not dispense water. To uncover other latent behaviour of the system, we applied our model to investigate the daily water withdrawal data, which covers the period of January to December 2018 (one year). There are a few other interesting observations that can be made from the plot in Figure 1. A few striking observations are:

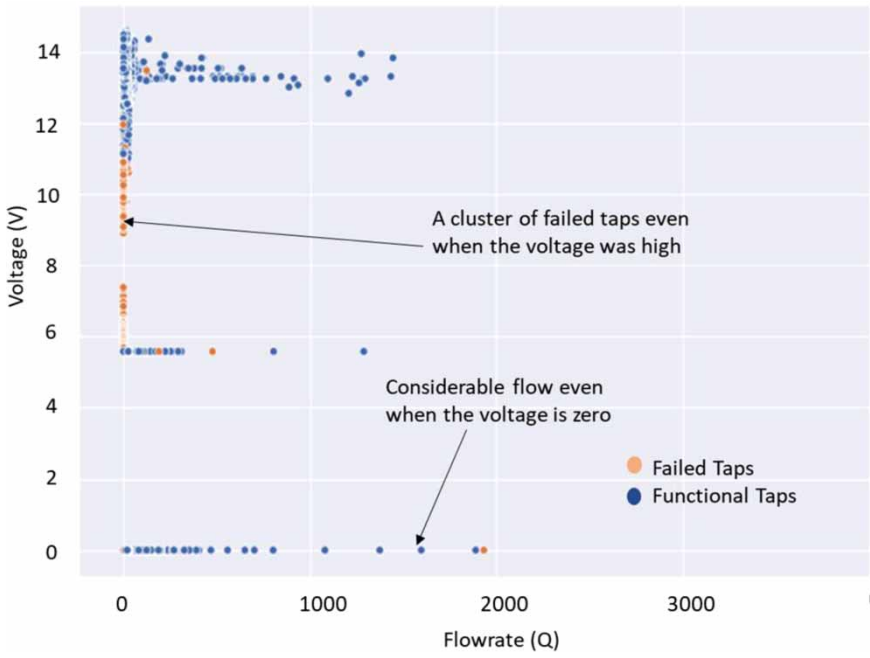

Figure 1 | Initial analysis of the dataset with a conventional method to show functionality and non-functionality of the EWSPs in the case study.

- Most of the water tap failures occur in the extreme right side of the plot, where *FlowRate* is the lowest. This may indicate a flow-related problem.
- There are two instances of failure where the *FlowRate* was recorded at the value of 1,900, but the voltage remained 0. This could be indicating a battery failure.
- There is one failure that occurred when the voltage was around 12.8 volts, but the *FlowRate* was around 15 litres, and it was recorded as a latent failure as well.
- There is another inspection that was performed to check if an error may be incorrectly indicating failures in the e-tap. The inspection shows that when the voltage was 0.00, there was still a reasonable quantity of flow (of up to about 13 litres). This may indicate a chipset error.

Based on the observations above, we coded some of the errors and used them as labels to investigate how they affect the sustainability of the electronic water withdrawal points. Table 2 below shows some of the sample errors and their associated codes (F1–F6).

Model development

As mentioned in the introduction section, we propose a novel predictive model based on the LSTM technique (Chollet 2018), which is capable of learning good representations of the available input time-series data. The representations can get closer to the expected output information for decision making.

After the preliminary investigation, we developed the proposed model (Figure 2) with Keras (one of the packages in the Python programming language that is widely used and supports LSTM modelling). Each of the LSTM memory cells has a 3D input (Wei *et al.* 2020). The choice of Keras in this study was because it gives room for consistency and provides useful feedback whenever there is a user error (Chollet 2018). The Keras application programming interface (API) is flexible and straightforward to use for a wide range of

Table 2 | Some of the errors investigated in the research

Failure definition	Failure code
No flow error	F1
Proxy error	F2
Voltage error	F3
Faulty valve error	F4
Master tag error	F5
Host valve error	F6

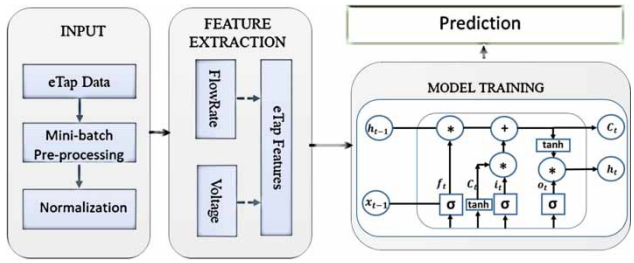

Figure 2 | The proposed model based on the LSTM network.

ML tasks. The LSTM layers were stacked so that the setup of the previous layer can be changed to give a 3D output. This was achieved by making the argument of the 'return sequence' on the last layer to True. The proposed model used for this study is a stacked LSTM network with eight hidden layers for 3D input in time processing. The hidden layers in the study are connected to a single output layer to produce a 2D output. The loss function used in the model design is the RMSE.

The central part of the proposed LSTM-based model is the cell state (which is the horizontal line running through the top in the 'MODEL TRAINING' box in Figure 2) down to the entire chain, with some linear interactions. The cells of the proposed model can remember values over a long period. The training model contains three gates that manage the flow of information in and out of the cell. The three gates in the model are: the input gate ($\boldsymbol{i_t}$), the forget ($\boldsymbol{f_i}$) gate and the output ($\boldsymbol{o_t}$) gate.

The input gate makes a decision on which new input to update to the current state of the cell; the forget gate determines the number of the previous states H_{t-1} that are allowed to go through the cell or what information to discard from memory. The output gate, on the other hand, makes the decision on the output based on the current state of the cell. The memory is then updated to hold the most current information based on the combination (or aggregation) of the old memory through the forget gate and the new memory state through the input gate.

These gates are introduced into the sigmoid function (σ) in Figure 2 to solve the vanishing gradient problem, and at time t it is computed as follows:

$$\boldsymbol{i_t} = \boldsymbol{\sigma}(\boldsymbol{W_i} * (\boldsymbol{W_h}, \boldsymbol{WH_{t-1}}) + \boldsymbol{b_i}) \quad (2)$$

$$\boldsymbol{f_t} = \boldsymbol{\sigma}(\boldsymbol{W_f} * (\boldsymbol{W_h}\,\boldsymbol{x_t}, \boldsymbol{WH_{t-1}}) + \boldsymbol{b_f}) \quad (3)$$

$$\boldsymbol{o_t} = \boldsymbol{\sigma}(\boldsymbol{W_o} * (\boldsymbol{W_h}\,\boldsymbol{x_t}, \boldsymbol{WH_{t-1}}) + \boldsymbol{b_o}) \quad (4)$$

$$Ç = \textbf{\textit{Softmax}}\ (W_c * (W_h x_t, WH_{t-1}) + b_c) \quad (5)$$

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (6)$$

$$H_t = o_t * \text{softmax}\ (C_t) \quad (7)$$

where

i_t and Ç – input gate values and the state of the memory cell at time t, respectively.
f_t – the value of the forget gate;
o_t – the output gate of the model;
C_t –the current state of the model;
W_i, W_f, W_o – randomly generated weight vectors for all three gates;
b_i, b_f, b_o – bias vectors for all three gates; and
H_t – the value of the memory cell at time t.

The sigmoid function (σ) serves to control the output between the values of 0 and 1 based on the current input of x and the prior output WH_{t-1}. At the computational phase of Equations (2)–(7), the weight and biases get trained by the model. This training is achieved by minimizing the loss between the outputs and the actual training samples of the proposed model.

Model implementation and validation

The proposed LSTM-based model is designed with a single input layer and eight LSTM-type hidden layers. A few other LSTM structures were examined as well, but the structure used for this study proved to perform better for the dataset. The number of neurons in the proposed model was initially set to 200 neurons, which was run against four hidden layers. However, the resulting predictions showed a terrible convergence of the model. Based on some of the limitations found in the dataset and the inconsistencies found in the time steps and date, the better option was to focus on every *AssetID* (tap) independently. Two different time steps that need to be programmed into the proposed model were considered. First is the *n_steps_in* parameter, which indicates the number of past time steps taken by the model to predict the corresponding output values (in time steps). The number of output time steps produced is dependent on the second parameter, which is *n_steps_out*. The final model includes the last 30 time steps from the dataset for different assets (taps) to predict future time steps.

The model fitness was validated using the technique called the walk-forward validation technique because of its ability to preserve temporal data (Falessi *et al.* 2018). In this case, the dataset gets split into two units (the training set and the testing set) that can be ordered. Then a decision is made on the minimum number of observations needed for the model training. This is done to obtain a configuration for the test setup. The training of the model then begins at the start of the time series with a design that enables it to make decisions for the next time step. Subsequently, the prediction is evaluated against the next time step in the time series. The predicted values were then measured using RMSE scores.

The hyper-parameters were tuned with the use of Talos, a library built for automated hyper-parameter tuning; it can be installed to work with Keras and it helped to reduce procedural redundancy.

The proposed model was built such that hyper-parameters will not be manually tuned. The accuracy of the model was another critical study criterion that was taken into consideration. For the model to be accepted as a useful prediction model, the accuracy has to be high enough. The proposed model was built on training sets to determine its accuracy; then, the testing sets were used as moderator samples to test the *trained* model. The prediction results were then compared with the actual values (ground truth) by calculating the accuracy based on Equation (8):

$$\text{accuracy} = \frac{\text{correctly predicted classes}}{\text{total testing class}} \times 100\% \quad (8)$$

While the Python programming language takes the weight of writing complex ML algorithms, we make a decision on the hyper-parameters to use in the model tuning, and this greatly influences the model's accuracy. Figure 3(a) and 3(b) show the results of model accuracy and loss from the experiment.

Figure 3(a) shows the proposed model's accuracy. The model's accuracy was measured using the RMSE, which gives the differences between the actual and the predicted values. Equation (9) below gives the formula used to calculate the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \quad (9)$$

where

N – number of total observations
x_i – actual value
$\hat{x}_i$ – predicted value

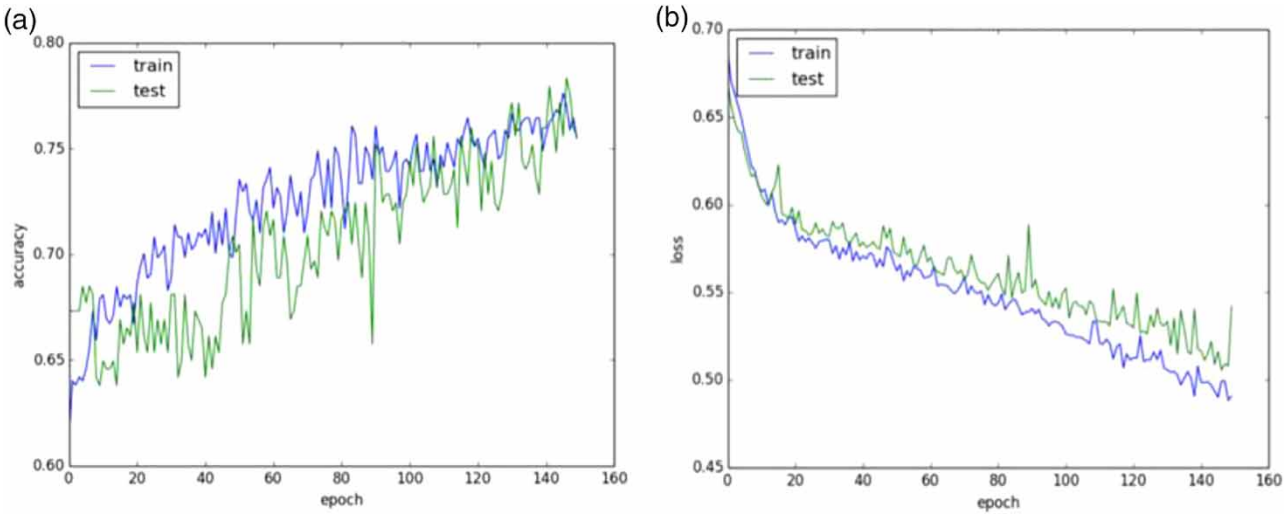

Figure 3 | Models showing accuracy and loss of the model. (a) Model accuracy, (b) Model loss.

The benefit of using RMSE is that it scales the scores and in the same unit as the values of forecast (i.e., hourly in our study). As seen from the diagram (Figure 3(a)), the accuracy of the model is increasing, with a current accuracy of 78%. The testing set shows this increment for categorical predictions only. So when different failures are categorized, the proposed model can easily predict the failure that is likely to occur in the future in the water taps. Figure 3(b) shows the training of the model's loss values for continuous predictions only. To make the model fit easily, we used fewer hyper-parameters. The epoch in the figure is the number of passes (times) that the proposed model's training vectors are used to update the weight.

Training the model is computationally very expensive. Our proposed model was built on a cloud-GPU platform, which hosts the Quadro P5000 graphics card. The time to complete a 100-epoch training (per tap) is, on average, 6 hours. Another GTX 1050 GPU was used for model training in parallel. It took roughly 8 hours to execute the task for one tap; this is because the program implementation needs a massive memory space to work effectively. With a cloud server, the training time of the model will reduce significantly.

RESULTS AND DISCUSSION

We proposed a predictive model for time-series prediction based on LSTM that is capable of predicting failures in rural water withdrawal taps. The model was built using the most fundamental principles and techniques, taking into consideration the many missing and incorrect values in the dataset (since the data is a very noisy dataset). The methodology presented in this paper is data-driven, and it is capable of self-learning to detect anomalies in the data. Concerning the above, the system can dynamically redesign itself to work for different time-series datasets.

The experimental model involves the use of a real dataset acquired from real solar-powered water taps. The data were preprocessed to remove impurities and further used as input for the prediction. The core of the LSTM model is the cell state and the three gates (the input, forget and the output gates as described in the 'Model development' section). The cell is the memory that holds information about the EWSPs. As the data are processed, information gets added or removed via the gates of the LSTM architecture. The gates, on the other hand, are different neural networks that are capable of making a decision about which information to allow into the memory or which one to forget during model training. Each of the gates has the activation function that scrunches values between 0 and 1 to help the cell state decide which data to allow or forget based on their importance. The model can categorize and predict failures for a real dataset and the model design can be further adjusted to suit the demand of future time series. A *complete* dataset would allow for higher quality predictions, accurately reflecting the future outcome for specific water tap assets that were analysed earlier. Two critical parameters that may be important to pay extra attention to would be: *n_steps_in* and *n_steps_out* (these parameters enable the model to take the past time steps *in* as a learning input, to produce predictions). The number of output

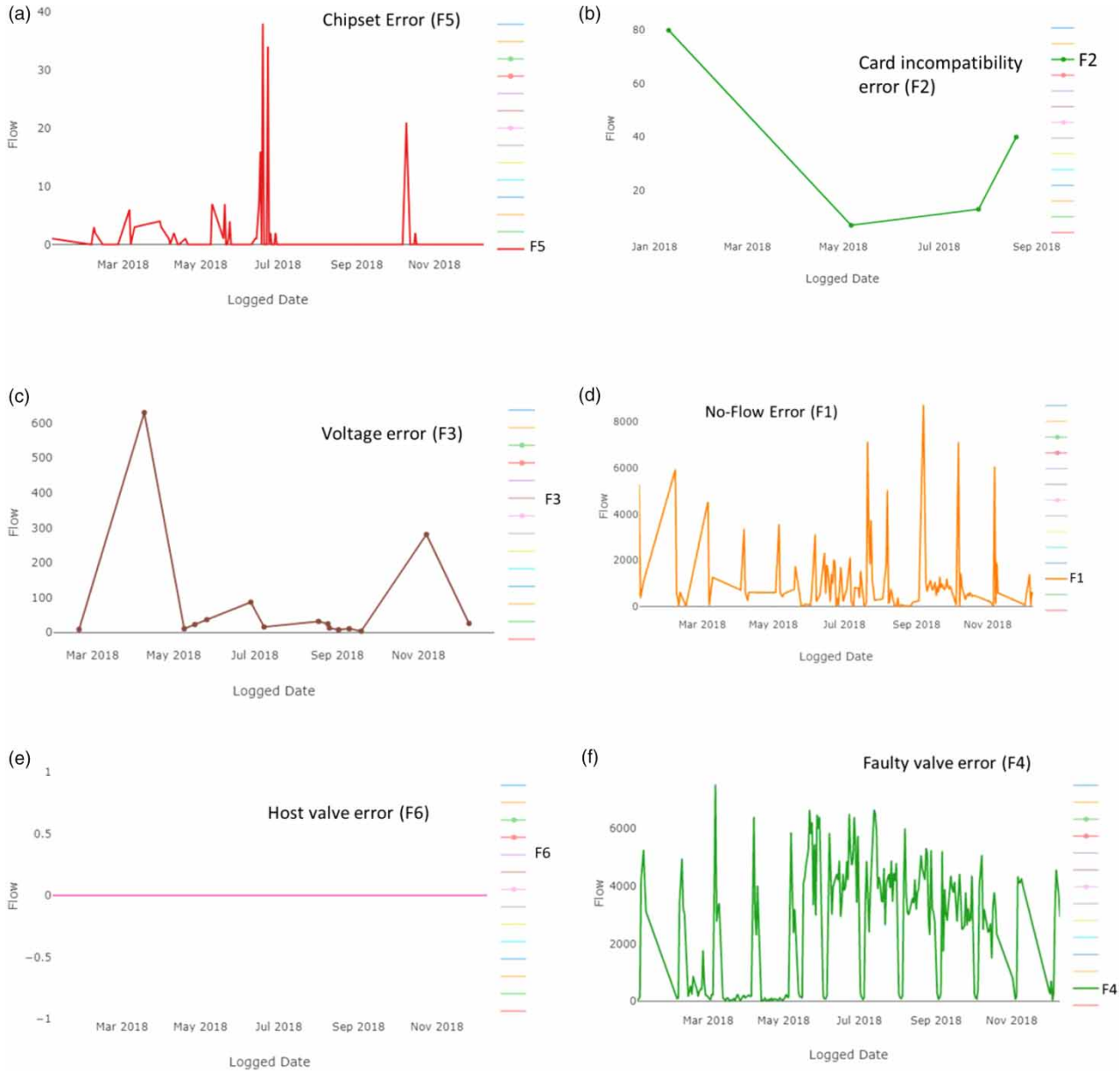

Figure 4 | Predicted errors/failures on some of the e-taps.

time-step predictions of the proposed model will depend on the value of *n_steps_out* (this parameter represents the number of time steps that are outputted as predictions).

The sample results for some of the failures are shown in Figure 4. Figure 4(a) shows a chipset error (F5), Figure 4(b) shows a card incompatibility error (F2), Figure 4(c) shows a voltage error (F3). Figure 4(d) represents error code F1, which is a flow-related error; Figure 4(e) shows a typical trend for failure type F6, which is a host valve error type. The figure shows that the error is predicted to remain at zero, which means that the error may not occur and cause the tap to fail. Figure 4(f), on the otherhand, shows a faulty valve error (F4). We observed that when there is a peak error, the tap stops functioning.

For all the results presented in this study, the two main empirical features used are the *Voltage* and *Flow-Rate* features, which were observed against the different failure trends in the *ErrorCode* column of the dataset. Figure 4(a) shows a monthly prediction of chipset error, which can cause the EWSPs to not dispense water. This kind of failure is caused when there is a technical mis-communication between a user token (pre-paid tag) and

an EWSP or a temporary malfunction of the EWSP. The proposed model is able to determine when the fault will occur and show its severity. Figure 4(b) represents a card incompatibility failure, which can occur when a token has been restricted or, mainly, because it was lost by a user. This failure can also arise due to the mechanical or technical malfunctioning of the EWSP. This kind of failure occurs but may not last for an extended period of time. Figure 4(c) captures the failure that can be caused by the voltage. When the voltage is below 6.0 volts, the EWSP will not have the capacity to power the e-tap. This failure can occur mostly during the rainy season, where there is less sunlight to charge the battery component of the EWSP system. Figure 4(d) shows the frequency of the failure caused by flow. Figure 4(e) shows a host valve error, and it can be noted that there was no error with the host valve throughout the period of study. Figure 4(f) shows a faulty valve error.

CONCLUSION

This research paper presents a model based on the LSTM paradigm for fault prediction in EWSPs in rural sub-Saharan Africa. It is actually the first time the method has been used to investigate e-taps. The proposed model takes into consideration two crucial pieces of input information from the solar-powered electronic taps, which are the flow rate and the voltage of the e-tap. The dataset used for the research is a real historical dataset with some inconsistencies.

From the result of the experiments, it has been shown that the LSTM model can discover latent information from the time-series dataset and can make informed predictions based on the results of the predictions. The proposed model's performance based on LSTM proves to have superior accuracy and efficiency for massive datasets (typically generated from daily use of the EWSPs in our case study) compared to the traditional statistical methods mentioned in the introductory section. The user of the model decides the best number of outputs needed. The size of these parameters influences the computation cost of the training of the proposed model.

From the results and the accuracy of the model, we conclude that the proposed model can provide reasonable predictions for the investigated tap failures. In the future, we hope to compare LSTM with other ML tools so as to generate a hybrid model for the development of an early-warning system for smart water taps.

ACKNOWLEDGEMENTS

This research has been achieved through a Ph.D. scholarship provided by TETFund Scholarship Grant (Nigeria), which is thankfully acknowledged. The authors would also like to thanks Rob Hygate, Roger Godwin and other members of staff of eWaterPay for providing access to the e-tap data.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Behailu, B. M., Hukka, J. J. & Katko, T. S. 2017 Service failures of rural water supply systems in Ethiopia and their policy implications. *Public Works Management & Policy* **22** (2), 179–196. https://doi.org/10.1177/1087724X16656190.

Cabrera, E. 2007 Statistical modeling for failure prediction in water supply networks. In *International Conference on Dependable and Quality Management*, June.

Chen, Y. & Wang, K. 2019 Prediction of satellite time series data based on long short term memory-autoregressive integrated moving average model (LSTM-ARIMA). In *2019 IEEE 4th International Conference on Signal and Image Processing, ICSIP 2019*. pp. 308–312. https://doi.org/10.1109/SIPROCESS.2019.8868350.

Chollet, F. 2018 *Deep Learning with Python*. Manning Publications Co., New York.

Falessi, D., Narayana, L., Fong, J. & Turhan, B. 2018 *Preserving Order of Data When Validating Defect Prediction Models*. pp. 1–20.

Foster, R. & Cota, A. 2014 Solar water pumping advances and comparative economics. *Energy Procedia*. (January). https://doi.org/10.1016/j.egypro.2014.10.134.

Foster, T., Willetts, J., Lane, M., Thomson, P., Katuva, J. & Hope, R. 2018 Risk factors associated with rural water supply failure: a 30-year retrospective study of handpumps on the south coast of Kenya. *Science of the Total Environment* **626**, 156–164. https://doi.org/10.1016/j.scitotenv.2017.12.302.

Fu, R., Zhang, Z. & Li, L. 2016 Using LSTM and GRU neural network method for traffic flow prediction. 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, pp. 324–328, doi: 10.1109/YAC.2016.7804912.

Jordan, M. I. & Mitchell, T. M. 2015 Machine learning: trends, perspectives, and prospects. *Science* **349** (6245), pp. 255–256.

Kini, B. V. & Sekhar, C. C. 2013 Large margin mixture of AR models for time series classification. *Applied Soft Computing Journal* **13** (1), 361–371. https://doi.org/10.1016/j.asoc.2012.08.027.

Laucelli, D. & Giustolisi, O. 2015 *Detecting Anomalies in Water Distribution Networks Using EPR Modelling Paradigm, (October 2016)*. https://doi.org/10.2166/hydro.2015.113.

Liu, P., Wang, J., Sangaiah, A. K., Xie, Y. & Yin, X. 2019 Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *MDPI – Sustainability* 1–14. https://doi.org/10.3390/su11072058.

Malhotra, P., Vig, L., Shroff, G. & Agarwal, P. 2015 Long short term memory networks for anomaly detection in time series. In *Artificial Neural Network, Computational Intelligence and Machine Learning*, April. pp. 22–24.

Mamun, M., Kim, J. J., Alam, M. A. & An, K. G. 2020 Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. *Water (Switzerland)* **12** (1). https://doi.org/10.3390/w12010030.

Mazumder, R. K., Salman, A. M., Li, Y. & Yu, X. 2019 Reliability analysis of water distribution systems using physical probabilistic pipe failure method. *Journal of Water Resources Planning and Management* **145** (2), 04018097. https://doi.org/10.1061/(asce)wr.1943-5452.0001034.

Mounce, S. R., Pedraza, C., Jackson, T., Linford, P. & Boxall, J. B. 2015 Cloud based machine learning approaches for leakage assessment and management in smart water networks. *Procedia Engineering* **119** (1), 43–52. https://doi.org/10.1016/j.proeng.2015.08.851.

Muharemi, F., Logofătu, D. & Leon, F. 2019 Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication* **3** (3), 294–307. https://doi.org/10.1080/24751839.2019.1565653.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R. & Muharemagic, E. 2015 *Deep Learning Applications and Challenges in Big Data Analytics*. pp. 1–21. https://doi.org/10.1186/s40537-014-0007-7.

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F. & Chandran, V. 2018 Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. *IEEE Access* **6**, 49325–49338. https://doi.org/10.1109/ACCESS.2018.2868361.

Patro, S. G. K. & Sahu, K. K. 2015 Normalization: a preprocessing stage. *International Advanced Research Journal in Science and Technology* **2** (3), 20–22. https://doi.org/10.17148/iarjset.2015.2305.

Ram, K. 2017 *Machine Learning Tools to Time Series Forecasting (December 2007)*. https://doi.org/10.1109/MICAI.2007.42.

Shewalkar, A., Nyavanandi, D. & Ludwig, S. A. 2019 Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research* **9** (4), 235–245. https://doi.org/10.2478/jaiscr-2019-0006.

Sina, D. & Thomas, B. 2019 Anomaly detection in univariate time series: an empirical comparison of machine learning algorithms. In *ICDM*. pp. 1–15.

Wei, X., Zhang, L., Yang, H. Q., Zhang, L. & Yao, Y. P. 2020 Machine learning for pore-water pressure time-series prediction: application of recurrent neural networks. *Geoscience Frontiers* (September 2019). https://doi.org/10.1016/j.gsf.2020.04.011.

Wu, Z. Y. & Rahman, A. 2017 Optimized deep learning framework for water distribution data-driven modeling. *Procedia Engineering* **186**, 261–268. https://doi.org/10.1016/j.proeng.2017.03.240.

Wu, Z. Y., El-Maghraby, M. & Pathak, S. 2015 Applications of deep learning for smart water networks. *Procedia Engineering* **119** (1), 479–485. https://doi.org/10.1016/j.proeng.2015.08.870.

Xu, Z., Ying, Z., Li, Y., He, B. & Chen, Y. 2020 Pressure prediction and abnormal working conditions detection of water supply network based on LSTM. *Water Supply* **20** (3), 963–974. https://doi.org/10.2166/ws.2020.013.

Yuanyuan, W., Ping, L., Wenze, S. & Xinchun, Y. 2017 A new framework on regional smart water. *Procedia Computer Science* **107** (Icict), 122–128. https://doi.org/10.1016/j.procs.2017.03.067.

First received 17 July 2020; accepted in revised form 26 October 2020. Available online 6 November 2020