A Deep Learning Approach to Predicting Household Water Usage in Karachi

Akeel Ather Medina
Computer Science (DSSE)
Habib University
Karachi, Pakistan
am05427@st.habib.edu.pk

Hana Ali Rashid
Computer Science (DSSE)
Habib University
Karachi, Pakistan
hr05940@st.habib.edu.pk

Hamna Jamil
Computer Science (DSSE)
Habib University
Karachi, Pakistan
hj05910@st.habib.edu.pk

***Abstract*—Water shortage is an increasingly critical problem for countries worldwide, and water usage prediction is crucial for effective long-term planning. This research presents a deep-learning approach to predict water usage for a household based on smart-meter data and weather conditions. We have used LSTMs for prediction based on existing literature for accurate long-time forecasting. The dataset used is the water usage data for a single household in Karachi that has been collected by the Karachi Water Project using smart meters. The goal of this project is to train a deep learning model to predict future water usage for the household, which can then be generalised for a larger number of households, and potentially predict water usage for the city of Karachi on a larger scale - and aid future water planning and distribution.**

***Index Terms*—Deep Learning, LSTM, Time-Series, Forecasting, Water**

I. Introduction

Pakistan has long been challenged by its constantly evolving settlements, especially in mega-cities like Karachi. This has led to challenges in infrastructure including water supply systems, to the extent that only 28% of dwellings in Karachi have reliable access to piped water supply [1]. This brings into stark contrast the dire need to not only understand the water usage in Karachi but also predict future usage for effective planning and distribution of such a scarce resource.

Smart meters have recently been part of a global effort to understand and predict water usage in households. They are installed at the primary water supply of each house and provide ”high-resolution” readings of water usage. These help water utilities in demand forecasting, regulating time-of-use watering, and making informed decisions in operations and planning [2]. The use of smart water meters is not only helpful in building better infrastructure but also in catching anomalies like non-revenue water and issues with infrastructure like leaking pipes [3]. Smart metering also provides improved billing for the consumers, as well as gives them insights about their water usage, and helps identify potential leaks in their households [2]. Smart devices like these have been developed and installed for study in selected households in Karachi as part of the Karachi Water Project [4]. We will be using the water usage data from one household over the time period of a year for the scope of this project.

The research question we aim to investigate is: How can deep learning be used to effectively predict household water usage in the context of Karachi, Pakistan?

This report starts with a systematic literature review of several aspects that must be considered during the course of this project, i.e. motivation for exploring water usage data, existing work for predicting household water usage, and deep learning for prediction of water usage in other contexts. We explore other sequence prediction studies for insights into architectures, models, and variables. We then discuss data acquisition and pre-processing, followed by the methodology implemented. The paper concludes with a discussion of results and future work.

II. Literature Review

Our initial goal for the project was to gain a holistic understanding of the landscape regarding water usage analysis, including traditional statistical analysis as well as more recent machine learning and deep learning-based techniques. This allowed us to identify variables to consider in our research, as well as guide us around challenges and limitations.

A. Extracting Household Behaviors from Water Usage

Researchers in [3] suggest two important aspects for analyzing residential water usage, analyzing water consumption trends in the overall household and water end usage within the house. The paper details their work on using machine learning to determine the appliances used in the household based on water usage patterns. There are several different technologies available in the market to detect flow and mechanical flow meters are most suitable for big scale due to their low cost. The data collected from flow meters is then analyzed for insights. Several researchers used supervised machine learning algorithms to detect appliances on the basis of flow rate, volume, etc. A challenge here is to separate concurrent events which are quite common in households and significantly affect the accuracy of insights so in order to deal with this, [5] used vector gradient technique to separate concurrent events [5]. Since this study only focused on water consumption trends, therefore, only flow rate and timestamp were used as data points. Household water use determines overall consumption during the month, week and day. To determine the appliance level use, all appliances/faucets were divided into two categories, automatic end-uses, and human-controlled end-uses. The disaggregation algorithm is used to separate data where the flow rate is higher than 0 for consecutive seconds. This disaggregated data is then fed to a k-means clustering algorithm to identify appliances. For identification, appliances were divided into four groups depending upon the two variables, average flow rate and duration [3]. This paper is an example of how water usage data can give a wealth of information about consumer behavior and how computational techniques like machine learning can allow even better analysis and prediction for water usage. A limitation of this paper was the simplistic model that took only two variables into account [3], which is something we can explore and possibly overcome in our work.

B. Household Water Usage Prediction

A research project [6] focuses on deep-learning for predicting household-level water usage in South Korea. They compared the performance of the traditional time series prediciton model ARIMA with LSTMs, considering the non-linear nature of human behavior and thus the non-linear water usage data. This project included water usage from a detached house, an apartment, a restaurant, and an elementary school, as representatives for each water usage type, and the data was collected across a year [6]. Both models were compared on the basis of Root-Mean-Square-Error (RMSE) and Correlation Coefficient (CC) [6]. The ARIMA model displayed an overfitting tendency.On the other hand, the LSTM model performed better overall as it considered external factors such as weather and weekdays or weekends on top of the water usage data [6]. This work is similar to the goals of our project, and is therefore a valuable resource for LSTMs, which we can consider in our work. However, the limitations of this model is that it can only predict water usage for no longer than the next day. It also did not consider national holidays, and has a short target area [6].

C. Deep Learning for Water Usage

A recent review paper [7] was highly informative regarding deep learning applications in managing water resources. Of the various fields of hydrology the paper details, we focused on the water resources management aspect. It shows that deep learning has been used in recent years for tasks including regression, classification, and sequence prediction. Of the 12 papers cited regarding water resources management, only two are said to have reproducible results [7]. LSTMs, GRUs, and CNNs were common in most of the projects, implemented mostly in Keras and TensorFlow [7]. DBNs and ANNs were also used in sequence prediction tasks [7]. The paper notes the lack of use of transformers for sequence prediction [7], which is something we can explore in our project. With respect to architectures, GRUs and LSTMs appear to have similar results but GRUs have fewer parameters and are simpler in terms of architecture and learning curve [7], which can also be a factor to consider in our work.

In another paper [8], researchers have investigated the correlation between climate and water usage in a district using a combination of CNN and Bi-LSTM in order to minimize the error. CNN model is generally efficient at predicting meteorological and historical water consumption while Bi-LSTM can predict long-term trends therefore the combination of two models is proven to be more accurate as compared to a single application of these models [8]. Correlation coefficient shows a strong correlation between the maximum temperature of the day and water consumption therefore max temperature is used as a temperature variable [8]. To predict water consumption, it was found that water consumption on the current day has a strong correlation with consumption of the previous five days [8]. Along with CNN and Bi-LSTM layer, Holiday and Temperature correction is also used whose coefficients are found after statistical analysis of previous data. The following formula is used

$$P1 = QP$$

where P is the predicted value by previous layer, $P1$ is the value after correction layer and Q is the coefficient. It was proved that using correction models can help improve prediction accuracy [8]. In our context, adding additional layers of correction could improve accuracy and make better predictions.

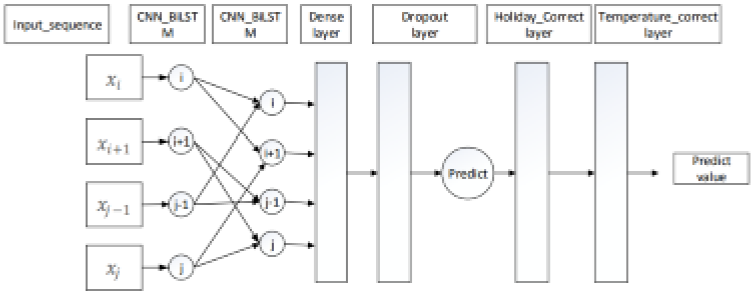


Fig. 1. Model architecture in [8]

D. Time-Series Forecasting using Deep Learning

Given the nature of our research question, time-series forecasting i.e. sequence prediction is the task we aim to implement in order to train a model to analyze patterns in water usage for our household data and predict future consumption.

A time series is a set of data collected at even intervals, and many deep-learning models have been developed to forecast a time series. A survey by Torres et al reviews Deep Learning architectures for this task, such as Recurrent Neural Networks (RNN), Long short-term memory (LSTM), and Convolutional Neural Networks (CNN). Torres et al identify the 3 main components of time series data:

- Trends, that is the main weight of the data, that can be described linearly, exponentially, etc.
- Seasonality, that is expected variations in the data at specific times.

- Residuals, that is fluctuations that make statistical analysis difficult.

As time series data consists of these 3 components, traditional statistical methods cannot accurately model it. Moreover, long-time series forecasting can be significantly harder using these methods [9]. Essien and Giannetti proposed an improvement for the traditional CNN and LSTM architecture focused on univariate time series forecasting using Convolutional LSTM Stacked Autoencoders, specifically tackling long-term time forecasting [10]. However, time series data may not be univariate, and a typical deep-learning model may not be effective in a correlated time series forecast. Thus, Wan et al proposed a Correlated Time Series Oriented LSTM (CTS-LSTM) that can capture patterns in correlated time series data [11].
Smart sensor data is often used as time-series data, and the proposed models can aid us in our research question. Specifically, the LSTM or CTS-LSTM model in prediction of household water usage in the context of Karachi, Pakistan.

III. Dataset

This project is in collaboration with KWP [4], and was thus granted access to data collected for their ongoing research projects.

Household water consumption can be measured in terms of the flow of water. Smart devices installed by the Karachi Water Project measured water flow rate and other variables as time series data of a single household over a year. The data collected contains the: flow rate, total flow, time received, and other data pertaining to the sensor's health. A complete list of variables measured and stored is given in Table I.

TABLE I
Raw Dataset For Household Water Flow

Variable	Description	Included
id	Unique sensor ID	Yes
node	Household name	No
time_sampled	UNIX time of sample	Yes
time_received	UNIX time of batch update	No
battery_level	Sensor battery level	No
flow_rate	Water flow since previous sample	Yes
temperature	Sensor internal temperature	Yes
flow_count	Water flow before unit conversion	No
signal_strength	Strength of communicating radio	No
total_flow	Cumulative volume of water used	Yes
total_flow_node	Cumulative flow per node	No

Some specifications of the smart meter device include,

- 30-second intervals between measurements
- Measures flow rate of water in Liters per minute
- Maximum sensor capacity is 60 Liters per minutes

IV. Methodology

A. Data Pre-processing

The dataset received included many columns, and for the scope of this study we did not require all of them. Therefore, we formed a new copy of the dataset and included the variables ”time_sampled”, ”flow_rate”, ”total_flow”, and ”temperature” variables from the raw data. Then, we added columns for datetime information as well as various time intervals to the data, for processing it for different intervals of time. We also added weather information in the ”avg_temp” column to include the average temperature of Karachi on the day of the measurement, to include that in our analysis. The columns of data in this extracted and expanded dataset are listed in Table II.

TABLE II
Modified Dataset For Household Water Flow

Variable	Description
id	Unique sensor ID
time_sampled	UNIX time of sample
flow_rate	Water flow in interval in Litres/minute
total_flow	Cumulative volume of water used
temperature	Sensor internal temperature
datetime	Converts time_sampled to timestamp
date	Just the date information from datetime
month	Stores the month in format YY-MM
time	Stores the timestamp from datetime
hour_count	Number of hours passed since first measurement
week_count	Number of weeks passed since first measurement
30min_count	Number of half-hour intervals passed since first measurement
5min_count	Number of 5-minute intervals passed since first measurement
volume	Volume of water in Liters used in measured interval
day	Day of the week of measurement
avg_temp	Average temperature of Karachi on the day of measurement

Next, we had to clean the data before using it to train a prediction model. For this, we visualized the data using Pandas and Matplotlib libraries for insights on its shape and possible anomalies. Figure 2 shows one such visualisation, and we see a large peak in both ”flow_rate” and ”volume” which reaches more than the sensor capacity of 60 liters per minutes. We also see a stretch where both variables are almost 0, which is another anomaly. We also see random small peaks i.e. fluctuation in the data which may also be noise. Moreover, another visualisation revealed that cumulative water volume used was not monotonically increasing as it should be, for the first two months of measurements.

Some types of anomalies that were flagged during this process include:

1) Sensor malfunction: flow_rate higher than sensor capacity.
2) Sensor malfunction: zero flow_rate when water was being used by the household.
3) Negative volume of water used in 30-seconds time interval due to possible mix-up of measurement sequence at server.
4) Noise and fluctuation: non-zero flow_rate when no water was being used by the household.
5) Sensor malfunction: missing measurements as flagged by a larger time difference between datapoints.
6) External factors: power outage, sensor battery depleted etc.

   As each anomaly was identified, we made new copies of the data without those data points. This is so as we want our initial prediction to be free from noise and anomalies

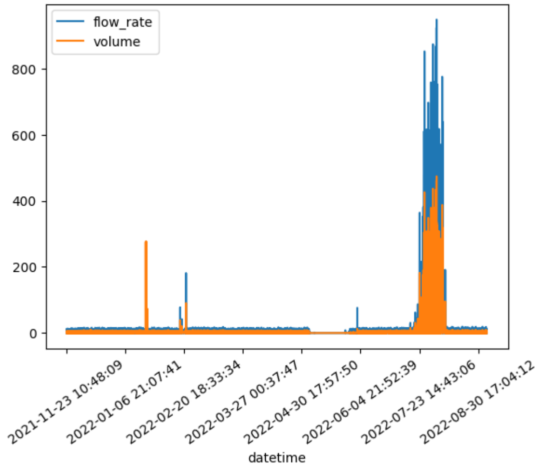


Fig. 2. Initial Data for water flow and volume.

would negatively impact prediction accuracy. However, this is not a cause of concern as a total of 5743 data points were removed, which still left a sizeable dataset of 712321 data points.

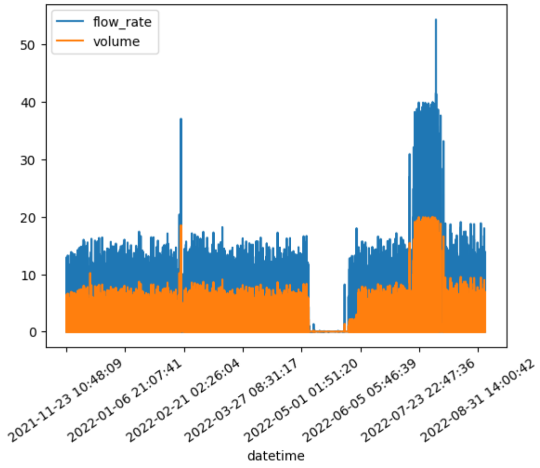


Fig. 3. Data with some anomalous data points removed.

B. Model and Architecture

A Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) that is capable of learning time-dependent sequences by using ”memory cells” within the network. These memory cells are responsible for remembering and predicting sequence data, such as hourly, daily, or weekly water usage. Thus, this type of network is well-suited for learning trends and seasonality of time-series data.

A basic LSTM architecture consists of four main components: an input gate, a forget gate, an output gate, and a memory cell. Each component is given its own vector of weights and bias values which are updated during training based on a forward and backward pass between each component and the input.

TABLE III
MODEL ARCHITECTURE

Name	**Layers**	**Units**	**Activation**
Bidirectional LSTM	1,3,6,8,11,13,15	256,128,64,64,64,64,64	Tanh
Dense Layer	16,17,18,19,20	40,30,20,10,1	Linear
Dropout	4,9	None	None
Batch Normalization	2,5,7,10,12,14	None	None

To forecast a time series, an LSTM network is first trained on past data sequences. A common technique in augmenting the data is to pass the data through a sliding window. The sliding window method takes each timestep of the time-series data and generates a new series with the past, current, and future timestep. After concatenating all results, the LSTM trains on this altered sequence. This allows the LSTM to learn a variable-length sequence that does not contain unnecessary historical data. In cases where the data is too large, a sliding window may help target certain trends and seasonality.

After training, the network is used to predict future values in the time series by taking into account the previous and current sequence. The output of the network is a prediction of a future sequence in the time series that is of user-defined length.

Based on our dataset and our goal for household water prediction, a simplified version of our model is presented in Figure 4. The model architecture is given in more detail in Table III.

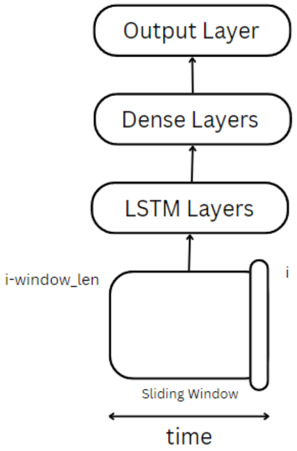


Fig. 4. Simplified LSTM

The layers of the model were chosen through trial and error, where this model gave the most accurate prediction on our dataset.

C. Implementation Details

Hyperparameter tuning is an essential part of deep learning models, as they configure a model for the dataset the model is used with. They are typically chosen through an experimentation process with trial and error. This is done by running the model with different hyperparameter configurations and evaluating the performance on a validation dataset. Common

TABLE IV
MODEL-SPECIFIC HYPERPARAMETERS

Model	Batch Size	Sequence Length	Window Stride	Epochs
5 Minutes	256	12	2	100
30 Minutes	20	24	6	100
1 Hour	10	24	6	100
1 Week	1	4	1	50

TABLE V
CONSTANT HYPERPARAMETERS

Hyperparameters	Value
Learning Rate	0.001
Optimizer	Adam
Dropout Rate	0.2
Loss Function	MSE
Batch Normalization	True

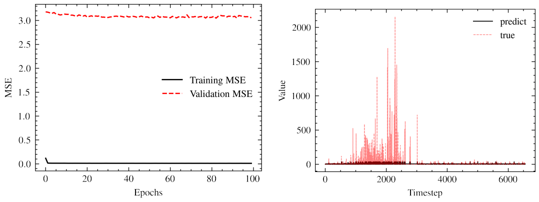
hyperparameters for an LSTM include the optimizer, learning rate, number of layers, size of the recurrent layers, type of activation functions, batch size, sequence length, and dropout rate.

Our dataset consisted of time-series data at 30-second intervals for an entire year, or for 10 months after anomalies were removed. To provide meaningful results while keeping in mind computational costs, we applied our LSTM model four times, while changing the timesteps of our data. In other words, we trained a model to predict water usage for the next 5 minutes, the next 30 minutes, the next hour, and the next week. A seperate model was also trained for the following versions of our data: raw data, cleaned data, and cleaned data with daily weather augmented to it.

Table IV depicts the hyperparameters that were changed to cater to a different input sequence for the different timesteps that we experimented with. The sequence length refers to the output of the sliding window, while the window stride refers to how many timesteps were skipped after construction of a window. The hyperparameters in Table IV-C were constant among all models.

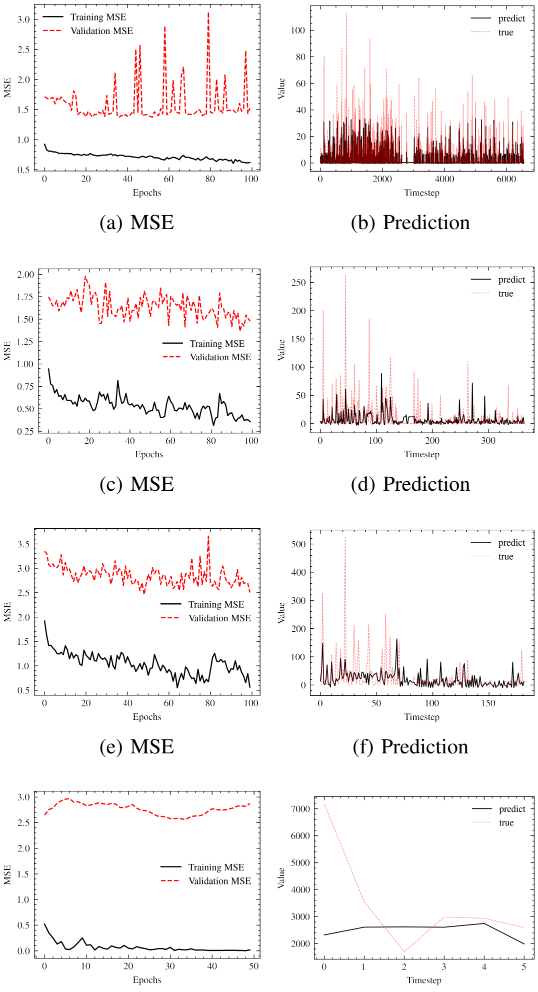
V. EXPERIMENTS

Training the models using the mentioned hyperparameters leads to the predictions shown in Figures 5, 6, and 7.



(a) MSE (b) Prediction

Fig. 5. Test Data prediction on Raw Data on 5 minute sequences



(a) MSE (b) Prediction

(c) MSE (d) Prediction

(e) MSE (f) Prediction

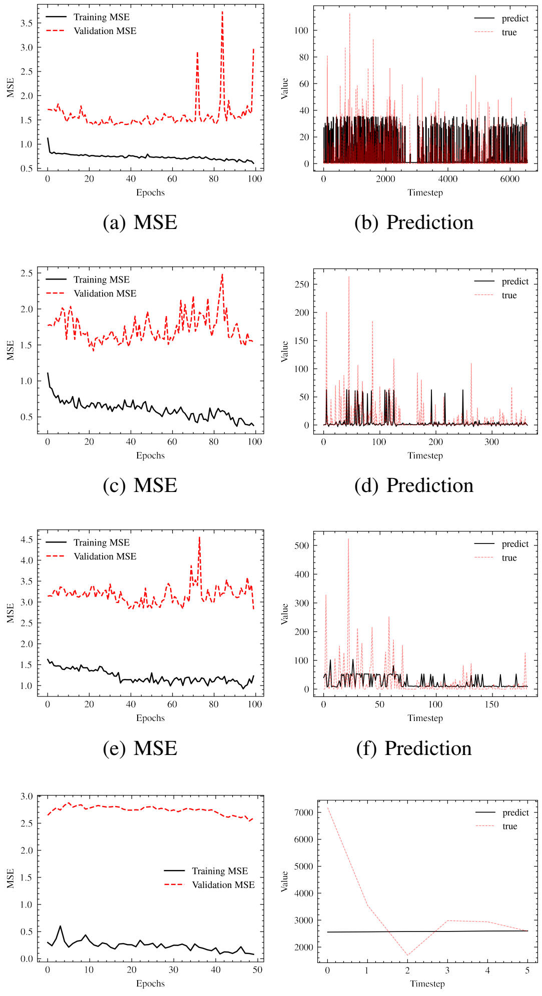
(g) MSE (h) Prediction

Fig. 6. Test Data prediction on Cleaned Data on 5 minute, 30 minute, 1 hour, and 1 week sequences

VI. DISCUSSION

A. Results

In terms of the models for usage prediction, we see that while there is some prediction, the models are not getting better on the validation data, that is they are either over-fitting on the training data, or the test data is from a different distribution than the train data. When looking at the predictions, we see an obvious increase in accuracy for data that was cleaned for anomalies, although weather data did not help much in the prediction. Moreover, we do see some obvious trends the model has captured and is able to reproduce, but seasonality and residuals are not really captured. Given the nature of this data, seasonality in terms of days of the week, seasons of the year, and even holidays and travel significantly impact water usage and are an important factor that must be included in any future prediction, especially for future planning. Therefore, more work needs to be done to include time information in the models and capture better trends in the models.



(a) MSE (b) Prediction

(c) MSE (d) Prediction

(e) MSE (f) Prediction

(g) MSE (h) Prediction

Fig. 7. Test Data prediction on Cleaned+Temperature Data on 5 minute, 30 minute, 1 hour, and 1 week sequences

B. *Challenges and Limitations*

One of the challenges faced in this study was the very raw form of the data. As this data was collected at high resolution for a year, it had a very large number of data points, which made it very tricky to distinguish noise and actual water usage. Due to seasonal changes in water usage, along with periods of no water use (such as travel), or higher-than-usual water use (such as with guests), and even calibration changes in the sensing device, there is a lot of variance in the data which requires more sophisticated statistical modelling to track. Therefore, this study is limited in its capacity of cleaning and pre-processing the data more than was done. Another challenge faced was the large number of parameters and hardware limitation in training models for prediction. This was made more challenging by the trial-and-error nature of constructing the model, due to the lack of existing models and datasets to use as benchmarks for predicting water usage, as also recognized by other researchers such as [7]. They mention the lack of detail about architectures and models by so-called ”Deep Learning” based studies, which have been shown to exploit the use of the keyword [7].

VII. Conclusion and Future Work

The scope of this study was to develop a pipeline in order to take raw data for water usage from sensor (smart meter) measurements, clean it for noise and anomalies, and feed it to a LSTM for predicting future usage. This report considered the many aspects that are related to predicting household water usage and looked at existing work among them. Based on the literature review, we used LSTMs as our priority in terms of model architecture. After acquiring our dataset, we pre-processed it to remove possible anomalies. Then, we kept the water flow rate and volume of water used along with time intervals as our primary variables for prediction. We also included weather information as that is observed to have an impact on water usage as well. Our models were trained on different versions of data, raw, cleaned, and cleaned with weather information. Cleaning the data increased prediction accuracy as expected, although weather information did not have much impact at this stage.

Our hypothesis at the beginning of the study was that Deep Learning can be effectively used to predict future household water usage for Karachi to be used in future planning. While our results cannot confirm this statement, we can conclude that there is potential here and with further data processing and model fine-tuning, we can have more promising results.

There is much potential for future work in this study, including pre-processing for anomaly detection and removal, as well as including time in the input. Moreover, the use of more specialized models such as LSTM-ALO or LSTM-PSO [7] may also contribute to better performance. There is also potential to study cyclic variation in days of the week, months, and seasons. Moreover, for more sophisticated analysis, this project has the potential of gaining behavioral insights from predictions such as holidays, seasons, time awake or asleep, or even the number of people in a household.

VIII. References

[1] V. A. Beard and D. Mitlin, “Water access in global south cities: The challenges of intermittency and affordability,” *World Development*, vol. 147, p. 105 625, Nov. 1, 2021, ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2021.105625. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305750X21002400 (visited on 10/17/2022).

[2] “Hedging for privacy in smart water meters - salomons - 2020 - water resources research - wiley online library.” (), [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020WR027917 (visited on 10/17/2022).

[3] G. M. Bethke, A. R. Cohen, and A. S. Stillwell, “Emerging investigator series: Disaggregating residential sector high-resolution smart water meter data into appliance end-uses with unsupervised machine learning,” *Environmental Science: Water Research amp; Technology*, vol. 7, no. 3, pp. 487–503, 2021. DOI: 10.1039/d0ew00724b.

[4] “Karachi water project.” (), [Online]. Available: https://www.karachiwaterproject.com/ (visited on 10/17/2022).

[5] K. Nguyen, R. Stewart, and H. Zhang, “An intelligent pattern recognition model to automate the categorisation of residential water end-use events,” *Environmental Modelling amp; Software*, vol. 47, pp. 108–127, 2013. DOI: 10.1016/j.envsoft.2013.05.002.

[6] J. Kim, H. Lee, M. Lee, H. Han, D. Kim, and H. S. Kim, “Development of a deep learning-based prediction model for water consumption at the household level,” *Water*, vol. 14, no. 9, p. 1512, Jan. 2022, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2073-4441. DOI: 10.3390/w14091512. [Online]. Available: https://www.mdpi.com/2073-4441/14/9/1512 (visited on 09/30/2022).

[7] M. Sit, B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir, “A comprehensive review of deep learning applications in hydrology and water resources,” *Water Science and Technology*, vol. 82, no. 12, pp. 2635–2670, Aug. 5, 2020, ISSN: 0273-1223. DOI: 10.2166/wst.2020.369. [Online]. Available: https://doi.org/10.2166/wst.2020.369 (visited on 09/30/2022).

[8] P. Hu, J. Tong, J. Wang, Y. Yang, and L. d. Oliveira Turci, “A hybrid model based on cnn and bi-lstm for urban water demand prediction,” *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2019. DOI: 10.1109/cec.2019.8790060.

[9] J. F. Torres, D. Hadjout, A. Sebaa, F. Martınez-Álvarez, and A. Troncoso, “Deep learning for time series forecasting: A survey,” *Big Data*, vol. 9, no. 1, pp. 3–21, 2021, PMID: 33275484. DOI: 10.1089/big.2020.0159. eprint: https://doi.org/10.1089/big.2020.0159. [Online]. Available: https://doi.org/10.1089/big.2020.0159.

[10] A. Essien and C. Giannetti, “A deep learning framework for univariate time series prediction using convolutional lstm stacked autoencoders,” in *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, 2019, pp. 1–6. DOI: 10.1109/INISTA.2019.8778417.

[11] H. Wan, S. Guo, K. Yin, X. Liang, and Y. Lin, “Cts-lstm: Lstm-based neural networks for correlatedtime series prediction,” *Knowledge-Based Systems*, vol. 191, p. 105 239, 2020, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2019.105239. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070511930557X.