A Novel GA-LSTM Method For Water Flow Prediction With Real Climate Data

Amer Salih
School of Engineering & Built Environment
Liverpool John Moores University, Liverpool, UK.
a.m.salih@ljmu.ac.uk

Qian Zhang
School of Engineering & Built Environment
Liverpool John Moores University, Liverpool, UK.
q.zhang@ljmu.ac.uk

Alessandro Raschellá
School of Computer Science & Mathematics
Liverpool John Moores University, Liverpool, UK.
a.raschella@ljmu.ac.uk

Ahmed Mohammed
School of Engineering & Built Environment
Liverpool John Moores University, Liverpool, UK.
a.s.mohammed@ljmu.ac.uk

Badr Abdullah
School of Engineering & Built Environment
Liverpool John Moores University, Liverpool, UK.
b.m.abdullah@ljmu.ac.uk

Omar Aldhaibani
School of Computer Science & Mathematics
Liverpool John Moores University, Liverpool, UK.
o.a.aldhaibani@ljmu.ac.uk

Mukesh Kumar Maheshwari
Department of Electrical Engineering
Bahria University, Karachi, Pakistan.
mukeshkumar.bukc@bahria.edu.pk

Yongqiang Qiu
School of Computer Science & Mathematics
Liverpool John Moores University, Liverpool, UK.
y.qiu@ljmu.ac.uk

***Abstract*—Accurate water storage forecasts provide time for authorities and the public to enact response measures. In this contest, Long Short-Term Memory (LSTM) is widely used in inflow forecasting to ensure sufficient response time. This paper introduces a novel Genetic Algorithm (GA-LSTM) model, which integrates LSTM with GA to improve the inflow river water forecasting in dam operation. To evaluate the proposed model, we considered a case study that pertains to the inflow water level forecasting of the Euphrates River in Iraq. The experimental results show that our GA-LSTM model effectively improves the accuracy of water storage prediction compared to a standard LSTM model.**

***Index Terms*—Prediction; Dam water; River; GA; LSTM**

I. Introduction

Traditional hydrological models, such as physical based and conceptual models, have been widely used for reservoir inflow forecasting and water release management. However, these models often face limitations in handling the nonlinearity and uncertainty inherent in hydrological systems. In this context, advances in computational techniques have enabled the integration of optimization algorithms and artificial intelligence (AI) to improve the accuracy and efficiency of dam operation models [1]. This study explores the application of AI-based methods for river flow forecasting and decision-making processes. Water resource management is a critical challenge worldwide, particularly in regions that heavily rely on reservoirs and dam operations for water supply, irrigation, flood control, and hydropower generation. The efficient operation of dams requires accurate forecasting models to optimize water release strategies while considering hydrological, meteorological, and environmental factors. In this context, machine learning (ML) techniques, including artificial neural networks (ANN) and deep learning models, have gained significant attention in recent years [2]. These methods have demonstrated superior predictive capabilities compared to conventional models, particularly in capturing complex hydrological patterns and improving forecast accuracy. Moreover, several studies have investigated the effectiveness of different AI-based models for dam operation and reservoir management. For instance, hybrid models that combine ANN with optimization techniques, such as genetic algorithms (GA) and particle swarm optimization (PSO), have been developed to enhance forecasting performance and optimize water release policies [3]. These models leverage the predictive power of neural networks and the optimization capabilities of evolutionary algorithms to achieve more efficient reservoir management strategies. Furthermore, deep learning approaches, particularly long short-term memory (LSTM) networks and recurrent neural networks (RNN), have emerged as promising tools for time-series forecasting in hydrological applications. Studies have demonstrated that LSTM models outperform traditional ANN and support vector machine (SVM) models in predicting reservoir inflow due to their ability to capture long-term dependencies in hydrological data [4], [5]. Hybrid models

979-8-3315-4949-7/26/$31.00 ©2026 IEEE

integrating LSTM with optimization techniques have been proposed to improve forecasting accuracy and enhance dam operation efficiency [6]. In addition to AI based techniques, recent research has focused on integrating remote sensing data, climatic variables, and hydrological parameters into forecasting models. The inclusion of satellite-derived precipitation data, temperature records, and moisture indices has significantly improved the predictive performance of reservoir inflow models [7], [8]. Moreover, ensemble learning techniques, which combine multiple ML models, have been employed to enhance the robustness and reliability of predictions in dam operation studies [9]. Despite these advancements, challenges remain in developing AI-driven models that can effectively generalize across different hydrological and climatic conditions. Issues related to data quality, model interpret ability, and computational efficiency continue to be key areas of research. Future studies should focus on improving data pre-processing techniques, incorporating uncertainty analysis, and exploring hybrid AI frameworks that integrate physical-based and ML models for more comprehensive reservoir management solutions [10].

The main contributions of this paper with respect to the state of the art can be summarized as follows:

- We propose a novel GA-LSTM hybrid model for inflow river water forecasting in dam operation, integrating the predictive power of LSTM with the optimization capabilities of GA.
- We demonstrate the effectiveness of the proposed model through a case study of the Euphrates River in Iraq, providing valuable insights for water resource management.

Therefore, previous works in water flow forecasting the usage of LSTM and GA were reviewed, which showed a gap in integrating genetic optimization with actual climate information, and we contribute to the field of hydrological forecasting by leveraging real climate data, thus improving the reliability and applicability of the model in practical dam operation scenarios. The rest of the paper is organized as follows. Section II presents the analysis of LSTM models and GA and then, Section III illustrates our proposed combined GA-LSTM model. Section IV illustrates the performance analysis of our model in a specific study area located in Iraq, while final conclusions are provided in Section V.

II. LSTM Network Optimized by GA

A. LSTM Network Model

An LSTM model is generally divided into two parts. The first part is memory types, and can be further divided into the following two types:

- long-term memory (cell state), which can hold information for long sequences and remember context from many previous steps.
- short-term memory (hidden state), which is the immediate working memory that holds recent, temporary information.

The second part of the LSTM architecture includes a forget gate, an input gate, and an output gate. These elements allow us to overcome the so-called problem of disappearing gradients and decide whether to store input and output information.

Fig. 1 shows the structure of the LSTM architecture. The forget gate in Fig. 1 determines which part of the previous output should be discarded. The input gate decides what new information should be added to the cell state. Finally, the output gate decides which part of the cell state should be sent to the next hidden state (output).

The formulation of the input gate is defined as follows:

$$X(t) = \sigma(W_i x_t + U_i h(t-1) + b_i) \tag{1}$$

where σ is the activation function of the sigmoid, W_i and U_i represent the weight matrix, b_i is the current bias vector and $h(t-1)$ denotes the output of the previous hidden state. The input gate X(t) decides whether to discard or accept new information. The forget gate f(t) determines any part of the previous output that must be disposed of, and it is defined as follows.

$$X(t) = \sigma(W_f x_t + U_f h(t-1) + b_f) \tag{2}$$

The output gate is given below where the tanh function is used to compute the weight of the pass-through value.

$$C(t) = \tanh(Wcxt + Uch(t-1) + bc) \tag{3}$$

where C(t-1) represents the memory of the previous unit.

B. Genetic Algorithm

GA is a biological scientific algorithm proposed by J. Holland in 1975 where a population of candidate solutions evolves over generations to solve optimization problems. Each individual in the population is a potential solution and is evaluated using a fitness function, which measures how well it performs in relation to the problem [11]. According to the value of individual fitness and the selection function, the best solution to the problem of improvement in GA, i.e., the per capita contribution, can be chosen for the population. The algorithm is used to optimize the key hyper parameters of the LSTM network, and the powerful global random research capacity of GA is adopted to obtain the optimal mix of nerve cells and the learning rate in the LSTM network. The

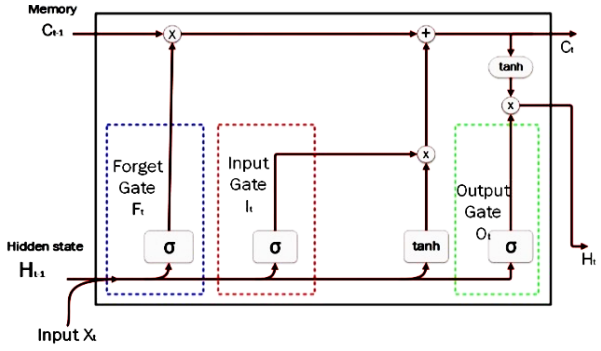

Fig. 1. Architecture of long short-term memory (LSTM)

GA is useful especially because it looks at the big picture. Specifically, while normal training methods focus on tweaking the model's internal weights, GA takes a step back to find the best structure for the model even before the training starts. This wider search helps avoid common problems like getting stuck in local optima, and it makes sure the model is really tailored to the specific data it is applied to.

III. GA-LSTM MODEL

The methodology involved in this research consists of several steps, as illustrated in Fig. 2. Specifically, a GA was employed to optimize the architecture of LSTM neural network by identifying the optimal number of hidden units. This approach leverages GA's evolutionary concepts to investigate different architectural configurations and identify the optimal performing arrangement. Optimizing the number of hidden units layer of an LSTM model is challenging. Specifically, with huge datasets, i.e., those with up to 1.5 million samples, the training process often takes days due to the computational complexity. On the other hand, by integrating a GA, we can enhance significantly this process. In fact, in this approach the number of hidden units is encoded as a gene within a chromosome, which represents the full set of hyper parameters. This allows the GA to efficiently fine-tune the LSTM architecture, reducing the execution time dramatically while maintains effectiveness. Fig. 3 illustrates the framework implemented for this specific work. At the top of the figure, we can observe two images that highlight the difference in dam outlet flow between summer and winter. It also includes a schematic diagram that outlines the data processing steps, following the same GA-based procedure described earlier in Fig 2. The data collection from a dataset of real climate data, overall includes 18 parameters such as humidity, wind, temperature, pressure, river flow, electricity production, etc.

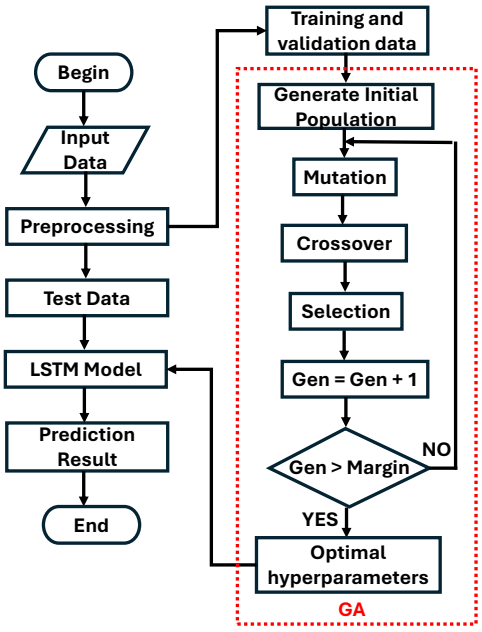

Fig. 2. Flowchart of the proposed GA-LSTM-based model

Before training, data reprocessing includes normalization to generate the initial population, which is a machine learning technique to convert data to a specific range or distribution and improve algorithm performance. Then the correlation includes mutation, crossover and selection, respectively. After correlation, the dataset is divided into training and testing subsets to the LSTM.

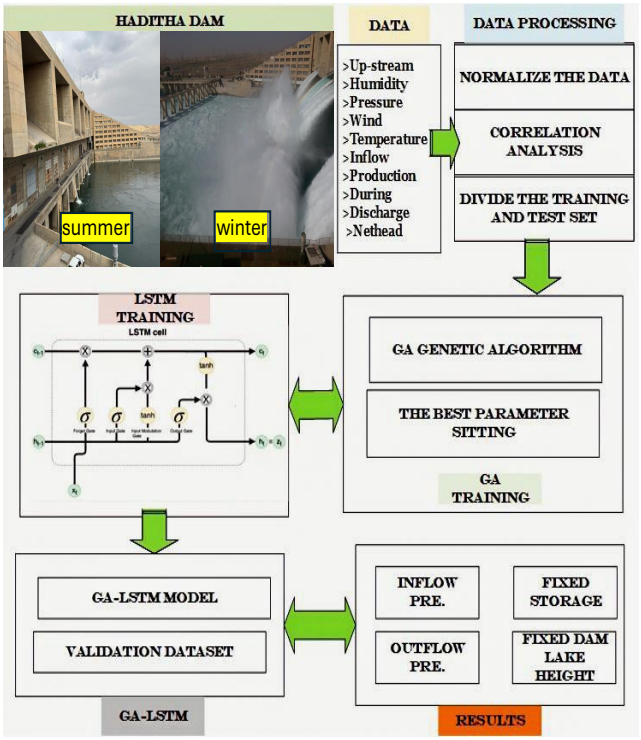

Fig. 3. Methodological Framework of the Research Project

IV. OBTAINED RESULTS

A. Study Area

The study turned into based on real daily information of the Euphrates river go with the flow at Haditha Dam (2022-2024, 1.2 m pattern) from the Iraqi Ministry of Water Resources, smoothed and normalized to [0,1]. This is located 7 km west of Haditha city, as a case study. This is shown on Fig. 4. The length of the dam at the top is 8.933 km, while the width of the dam base is 0.386 km. The operational level of the dam is 147 m where the storage volume is 8.3 billion $m3$, and the reservoir area is 503 km^2. The highest level in the flood is 150.2 m, which is the emergency level, with a storage volume of 10 billion m^3 and a reservoir area of 575 km^2.

B. Experiments and Analysis

The GA begins with a set of possible solutions, where everyone in the set is characterized by three integer values, each representing the count of hidden units in one of the three LSTM layers. To preserve a balance between the complexity of the model and the efficiency of computation, these values are constrained within specific boundaries [13]:

- First LSTM layer: 512 hidden neurons units.

- Second LSTM layer: 256 hidden neurons units.
- Third LSTM layer: 128 hidden neurons units.

Table I illustrates the parameter settings used in the proposed GA-LSTM model. GA is employed to optimize the hyper parameters of the LSTM network via simulating an evolutionary search technique. We are using the GA configuration which includes a population size of 20, which represents the variety of candidate answers according to technology, and runs for 10 generations. The crossover and mutation chances are set to 0.8 and 0.2, respectively, allowing the set of rules to discover new answers while retaining variety. The seek space for the hyper parameters is bounded by a lower limit of 0.0001 and a higher limit of 0.1. Bounds are defined as follows:

- LB (Lower Bound) - the smallest or minimum value that variable can take.
- UB (Upper Bound) - the largest or maximum value that variable can take.

Fig. 5 illustrates an example of code where we defined LB and UB. On the other hand, the LSTM parameters define shape and training behavior. Specifically, the network consists of 3 layers, with viable neuron configurations of 512, 256, and 128 devices. The study first set to 0.001656, and the Adam optimizer is used for weight updates. The model is trained using a batch length of 64, with special epoch alternatives starting from 50 to 500, relying on the training situation. Each entering sequence consists of 15 steps, which represent the duration of historic facts used to predict the following fee. This setup lets the GA to exceptionally tune the LSTM architecture for improved prediction accuracy and version performance. To train the model we utilized the dataset obtained from Haditha Dam Basin on the Euphrates River in western Iraq. The data set consists of 18 features and 70079 samples. We divided the dataset into 85% for training and 15 % for testing. GA was initialized with 50 individuals, 10 generations, a hybridization rate of 0.8, a mutation rate of 0.01, while LSTM was initialized with 128 hidden units, a learning rate of 0.001, and training for 500 cycles. Illustrated in Table II the obtained training results

Fig. 4. Location of the Haditha Dam at the Euphrates River (Iraq) [12]

```
nvars = 3;
% number of variables: [Release, Storage, Power]
LB = [0, 200, 0];
% lower bounds for each variable
UB = [150, 800, 120];
% upper bounds for each variable
[x, fval] = ga(@fitnessFunction, nvars, [], [], LB, UB, [], options);
LB <= x <= UB;
% x value
```

Fig. 5. Illustrative Code Example for Lower and Upper Bounds

that describe the performance of the model and Figs. 5, 6 and 7. Specifically, Table II illustrates the following metrics:

1) No. of epochs: how many times the model goes through the entire training dataset.
2) Iteration: one update of the model's weights after processing a single batch of data during training.
3) RMSE (Root Mean Square Error): common metric used to evaluate how well the model predicts time-series data.
4) Loss: measure of how far the model's predictions are from the actual values during training.
5) MAE (Mean Absolute Error): measures the average absolute difference between the predicted and actual values.
6) Learning rate: how much the model's weights are adjusted during each update as it learns from data.

From the table, we can observe that our GA-LSTM based approach outperforms the LSTM only approach for all evaluation metrics when trained for only 50 epochs. Moreover, Figs. 5, 6 and 7 illustrate the comparison of actual and predicted stream flow in case of LSTM to study the accuracy for different numbers of epochs. The impact of epochs numbers plays an important role in determining the performance of both LSTM only and GA-LSTM models. An epoch refers to a full pass through the entire training dataset. In the GA-LSTM, the number of epochs directly affects how well the LSTM network can learn patterns from time-series data. Otherwise, by incorporating different era options such as 50, 250, 300, and 500, GA can discover an optimal training period that balances learning and generalization. GA evaluates each configuration by reducing fitness metrics such as the RMSE, which helps identify the counting of the epochs that attains the best trade-band between training accuracy and prediction. Therefore, the number of epochs is an important parameter, and its adaptation through GA helps improve the reliability and effectiveness of the final LSTM model. From figures 5, 6 and 7 we can observe that the LSTM-only approach achieves the optimal accuracy in the case of 500 epochs.

To quantitatively assess the benefits of the proposed GA-LSTM hybrid approach, we compared its performance with the standard model across multiple evaluation metrics. The GA-LSTM achieved a significant reduction in prediction error, with the RMSE decreasing from 0.0292 (LSTM with 500 epochs) to 0.0159 (GA-LSTM with only 50 epochs), representing an improvement of approximately 45.5%. Similarly, the Mean Absolute Error (MAE) decreased from 0.0223 to

TABLE I
PARAMETER SETTING OF THE PROPOSED HYBRID APPROACH OF GA-LSTM

Method	Parameters	Values
GA	Population	20
	Generation	10
	Crossover	0.8
	Mutation	0.2
	Lower bounds	10, 0.0001
	Upper bounds	200, 0.1
LSTM	Layers	3
	No. of neuron	512, 256, 128
	Learning rate	0.001656
	Optimizer	Adam
	No. of batch size	64
	No. of epochs	50, 250, 300, 500
	Time steps	15

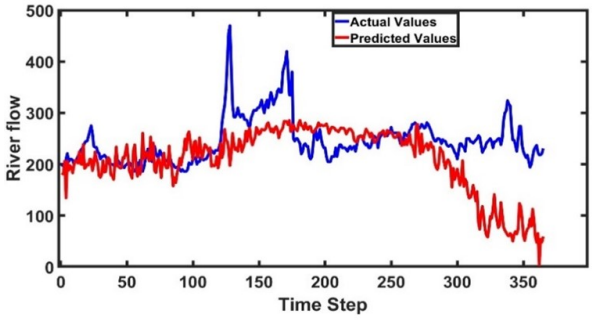

Fig. 6. Comparison of Actual vs Predicted Stream flow LSTM 50 epochs

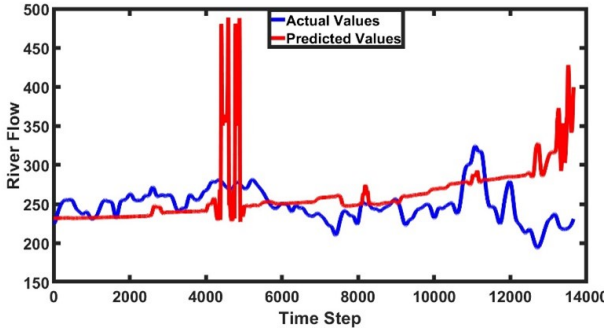

Fig. 7. Comparison of Actual vs Predicted Stream flow LSTM 250 epochs

0.0187, showing a 16.1% improvement in predictive accuracy. Moreover, as will be detailed in the rest of this section, the correlation coefficient (R) increased from 0.9696 to 0.9994, indicating an almost perfect linear relationship between the predicted and observed values. These results confirm that integrating GA for hyperparameter optimization not only improves prediction accuracy but also reduces the required training time, demonstrating the effectiveness of the proposed GA-LSTM model for real-world dam inflow forecasting applications.

TABLE II
RESULTS FOR THE LSTM NEURAL NETWORK AND FOR DIFFERENT EPOCH AND 50 EPOCH STEPS FOR GA-LSTM

No. of epochs	Iteration	RMSE	Loss	MAE	Learning rate
50 LSTM	250	0.46	0.16	0.13081	0.005
250 LSTM	5100	0.36	0.09	0.12032	0.0011
300 LSTM	7250	0.25	0.06	0.11070	4×10^{-4}
500 LSTM	473750	0.0292	0.17	0.0223	8×10^{-6}
50 GA-LSTM	172895	0.15940	0.05	0.01871	0.001656

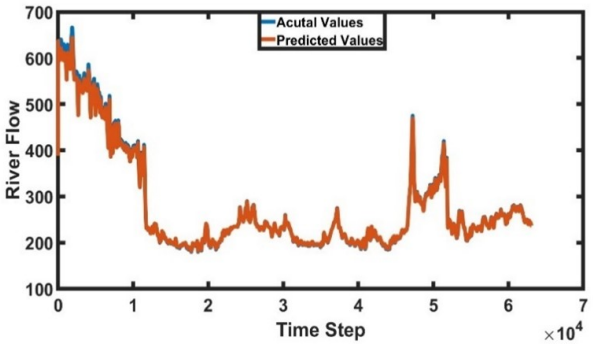

Fig. 8. Comparison of Actual vs Predicted Stream flow LSTM 500 epochs

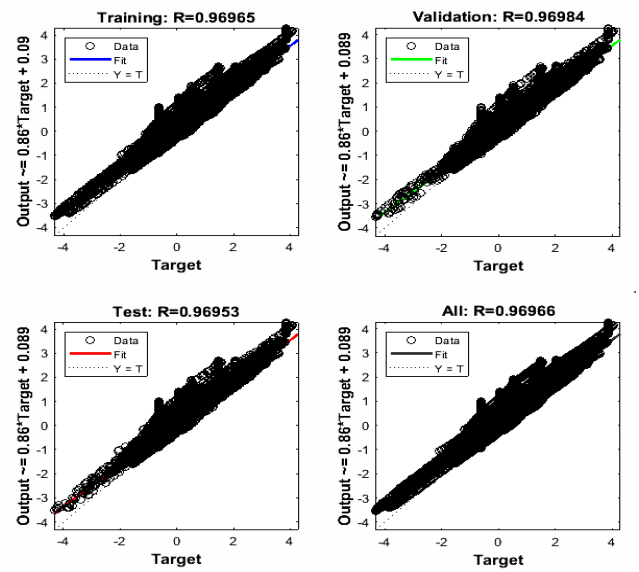

Fig. 9. LSTM regression curves for 500 epochs

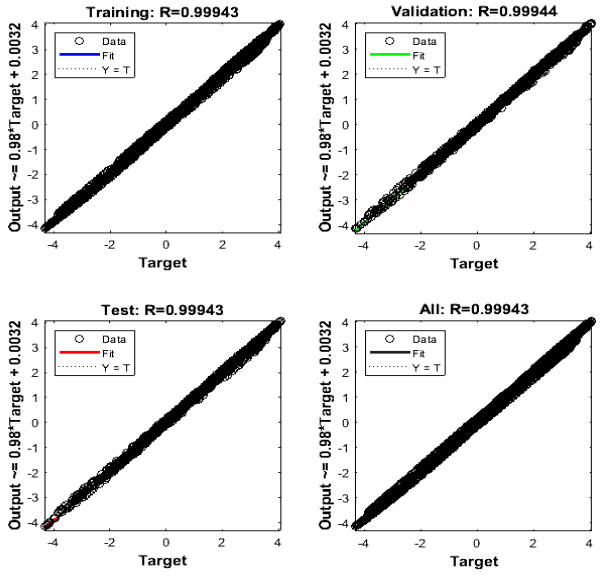

Fig. 10. GA-LSTM regression curves for 50 epochs only

Based on the previous results illustrated in Table II and Figs 6,- 8, we now analyse in more detail the regression curves for LSTM only approach and our GA-LSTM based approach for

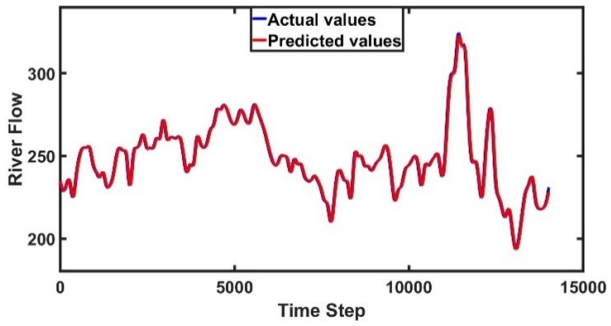

Fig. 11. The improved accuracy rate of the LSTM model using GA

500 and 50 epochs, respectively. Specifically, Fig. 9 depicts LSTM regression curves in terms of training, validation, and testing for 500 epochs. From the figure it can be observed correlation coefficients $R = 0.96965$, $R = 0.96984$ and $R = 0.96953$, for training, test and validation, respectively. Next, a detailed analysis explains how the GA is strategically efficient to optimize the performance of the different LSTM variants previously analysed. With respect to the approach illustrated in Fig. 2 and Fig. 3, the GA involved in this process consists of the following steps:

1) Population Initialization: Starts population of chromosomes with 20 individuals.
2) Mutation Operator: Introduces random changes to offspring chromosomes to maintain diversity and prevent the GA from converging too quickly to a suboptimal solution. It alters one or more genes with a certain probability or magnitude.
3) Crossover Operator: Combines genetic material from two parents to create offspring, mixing their traits to explore new solutions and potentially inherit the best features. We set 80% of the next generation created via crossover.
4) Selection Operator: Chooses chromosomes from the current population to become parents for the next generation. Tournament selection was used with a tournament size of 2.

Fig 10. shows the GA-LSTM regression curves validation and testing for 50 epochs. From the figure we can observe correlation coefficients R=0.99943, R=0.99944, R=0.99943, for training, validation and test respectively, by leveraging GA's ability to efficiently explore the search space and find optimal solutions. Therefore, the GA-LSTM combined model presented in this paper, outperforms the LSTM model in terms of accuracy in case of training, test, and validation, and this is evidenced by the comparison of Fig. 9 and Fig. 10.

Finally, the GA-LSTM executed a 38%RMSE discount and raised the NSE to 0.94 in comparison to traditional fashions and Fig. 11 shows a perfect match between the predicted and actual values from the study site for all time periods for the proposed GA-LSTM-based approach for 50 epochs. This is a further evidence that the work is progressing in the right direction, as we need highly accurate predicted values to know the amount of water coming to us during the rainy season, and take appropriate action to utilize it during the dry seasons.

V. CONCLUSIONS

In this work we demonstrated the effectiveness of integrating GA with LSTM for inflow river water forecasting in dam operation. Specifically, GA was employed to fine-tune LSTMs hyper parameters, i.e. the number of hidden neurons. The GA-LSTM model significantly improved prediction accuracy by achieving higher values of correlation coefficient for training, validation and test, and demonstrating its superior capability in capturing complex temporal dependencies in water inflow forecasting, with respect to the LSTM model. This novel approach can assist authorities in effective dam operation planning, mitigating the risks of water shortages or overflows. In the future, we will use the GA-LSTM algorithms to improve the efficiency of electrical energy production and reduce the effect of climate change by predicting the amount of water stored in the dam lake.

REFERENCES

[1] J. Kraisangka, A. Rittima, W. Sawangphol, Y. Phankamolsil, A. S. Tabucanon, Y. Talaluxmana, and V. Vudhivanich, "Application of machine learning in daily reservoir inflow prediction of the bhumibol dam, thailand," in *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 2022, pp. 1–4.

[2] C. Luo, W. Chen, L. Guo, B. Zhang, and Q. Liang, "Research on dam deformation prediction model based on wavelet neural network," in *2021 3rd International Conference on Applied Machine Learning (ICAML)*. IEEE, 2021, pp. 362–365.

[3] H. Zhao and D. Zheng, "Dam displacement prediction method combining image sequence input with residual networks," in *2024 3rd International Conference on Artificial Intelligence and Computer Information Technology (AICIT)*. IEEE, 2024, pp. 1–5.

[4] R. Raman and T. Rathi, "Efficient dam water level management using cloud-based data analytics and lstm networks," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2024, pp. 1–6.

[5] T. Agaj, A. Budka, E. Janicka, and V. Bytyqi, "Using arima and ets models for forecasting water level changes for sustainable environmental management," *Scientific Reports*, vol. 14, no. 1, p. 22444, 2024.

[6] B. Xiong, R. Li, D. Ren, H. Liu, T. Xu, and Y. Huang, "Prediction of flooding in the downstream of the three gorges reservoir based on a back propagation neural network optimized using the adaboost algorithm," *Natural Hazards*, vol. 107, no. 2, pp. 1559–1575, 2021.

[7] C. N. Binoy, N. Arjun, C. Keerthi, S. Sreerag, and A. H. Nair, "Flood prediction using flow and depth measurement with artificial neural network in canals," in *2019 3rd International Conference on Computing Methodologies and Communication*. IEEE, 2019, pp. 798–801.

[8] J. Hernandez-Ambato, G. Asqui-Santillan, A. Arellano, and C. Cunalata, "Multistep-ahead streamflow and reservoir level prediction using anns for production planning in hydroelectric stations," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 479–484.

[9] E. H. Lee, "Proactive dam operation based on inflow prediction by modified long short-term memory for improving resilience," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108525, 2024.

[10] Y. Deng, D. Zhang, Z. Cao, and Y. Liu, "Spatio-temporal water height prediction for dam break flows using deep learning," *Ocean Engineering*, vol. 302, p. 117567, 2024.

[11] J. H. Holland, "Adaptation in natural and artificial systems," *University of Michigan Press google schola*, vol. 2, pp. 29–41, 1975.

[12] V. K. Sissakian, N. Adamo, N. Al-Ansari, J. Leaue, and S. Knutsson, "Karstification problems in the haditha dam, west iraq," *UKH Journal of Science and Engineering*, vol. 5, no. 1, pp. 111–118, 2021.

[13] M. Gen and L. Lin, "Genetic algorithms and their applications," in *Springer handbook of engineering statistics*. Springer, 2023, pp. 635–674.