https://doi.org/10.2352/EI.2023.35.3.MOBMU-367

Data-Driven Approach for Robust Flood Prediction

Ganesh Reddy Gunnam*[1] *, Devasena Inupakutika*[1] *, Rahul Mundlamuri*[1] *, David Akopian; The University of Texas at San Antonio; San Antonio, Texas

Abstract

Time-series prediction problems have been effectively solved by deep neural networks lately given their ability to understand temporal characteristics found in time series. In this study, a deep learning-based flood occurrence prediction method, LSTM-PCA is presented for the successful interpretation of weather events and meteorological data with higher accuracy. The proposed model is evaluated on the United States National Climate Data Center (NCDC) dataset, and NCDC storm events. Correlation analysis was performed on the meteorological and weather phenomenal data for choosing the appropriate parameters. The experimental results show that the model achieved 96.49% accuracy while predicting floods in the United States from the year 2013 to 2019.

Introduction

Flooding is one of the most destructive natural disasters in the world, causing significant damage to infrastructure, homes, and lives. Early warning and prediction of floods are essential for reducing their impact [1]. various data sources, including meteorological, hydrological, and topographical data. Meteorological data includes information on precipitation, temperature, and wind, while hydrological data includes information on river flow, water levels, and soil moisture. Topographical data includes information on the terrain and land use, such as elevation and land cover. In addition to these data sources, other information that can be useful for flood prediction includes historical flood data, land use and land cover maps, and data from remote sensing devices such as radar and satellites [2]. However, due to the changing nature of climatic conditions, it is essentially difficult to estimate the time and location of floods. As a result, the main flood prediction models used today are primarily data-specific and rely on a number of simplified assumptions [3]. In recent years, machine learning and deep learning has emerged as a powerful tool for making predictions in various fields, including weather forecasting and hydrological modeling [4]. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are commonly used to analyze this data and make predictions. These models can be trained on large amounts of historical data to learn patterns and make predictions [5]. The combination of various data sources and deep learning models can provide an accurate and robust approach for flood prediction, enabling early warning and mitigation of flood impacts. Furthermore, because of the accessibility of vast amounts of meteorological data and the computational effectiveness of deep learning techniques, it may be used to predict floods and provide outcomes comparable to those of conventional models. Deep learning models are a good substitute for sophisticated physical models in the flood prediction process because they do not require a complete grasp of the physical system that explains the atmosphere surrounding the globe [6]. For training and testing proposed deep neural networks we have used a standard large set of realtime open source data which is available from United Sates National Climate Data Center (NCDC) and National Oceanic and Atmospheric Administration (NOAA).

The specific contributions of this work are as follows:

1. First, correlation analysis was performed as part of exploratory data analysis on the meteorological and weather phenomenal data for choosing the appropriate parameters.

2. Second, a deep learning-based flood occurrence prediction method called LSTM-PCA is presented for the successful interpretation of weather events and meteorological data with higher accuracy.

3. Finally, a comparative performance evaluation of LSTM-PCA model was performed with the United States National Climate Data Center (NCDC) dataset, NCDC storm events with SVM (machine learning) and LSTM methods.

The remainder of this paper is organized as follows. Section II presents experimental setup for dataset details and data preprocessing techniques. Section III discusses the proposed methods SVM and LSTM with PCA. In Section IV, results are discussed. Conclusion and future work are present in section V.

Related Work

There has been a significant amount of research in the field of using deep learning for flood prediction. Some previous works have used various types of neural networks, such as recurrent neural networks (RNNs) [7] and convolutional neural networks (CNNs) [8], to analyze historical flood data and make predictions about future floods. These studies have been able to achieve high levels of accuracy in their predictions, and have also been able to identify important factors that contribute to flood risk. Other works have used deep learning to process satellite imagery to detect and predict floods. Some works have also used deep learning to predict floods by analyzing social media data to detect early warning signs of floods [9]. In terms of improved deep learning-based models, [10] proposed an improved long-term short-term memory neural network LSTM flood forecasting model, for stream-flow prediction implemented a deep learning model for small watershed stream flow forecasting based on LSTM. The dataset considers past stream-flow data, past weather data, weather forecast data of the hydrological stations.

The work in [11] presented one machine learning model to predict floods and send that alert by email. They have combined three different datasets together by using GeoPandas, GeoJSON

[1]These authors contributed equally.

with the help of longitude, latitude and time. These datasets are from United Sates NCDC and NOAA. The datasets used in their research were named NCDC Storm events dataset, NOAA Daily summaries dataset and NOAA precipitation Reconstruction dataset. In the NCDC storm dataset a field named event_type, which is type of event such as Flood, Flash Flood, Hail, Thunderstorm wind, Heavy Rain etc., Out of these events Flood and Flash Flood considered as Flood which is labeled as 1 and all others considered as no flood which is labeled as 0. They feed this dataset to Random forest machine learning algorithm. They achieved 80% of accuracy, 85% of precision, 80% of recall and 82% of F1 score.

Dataset details and Preprocessing

The National Centers for Environmental Information (NCEI) is a division of the National Oceanic and Atmospheric Administration (NOAA) that maintains and provides access to a wide range of climate-related data sets, including historical weather and climate observations, as well as climate model output and projections. The NCEI's Climate Data Online (CDO) portal provides access to such data sets, which can be used for a variety of research and application purposes, including climate prediction.The NCEI maintains a large archive of climate-related data that can be used to train and evaluate deep learning models for climate prediction. Some relevant examples of data sets include Historical weather observations, such as temperature, precipitation, and wind data, which can be used to train models to make local and regional weather forecasts. Climate reanalysis data, which combines historical observations with numerical models to provide a consistent, high-resolution view of the climate system over the past several decades. Climate model output, such as from the Coupled Model Intercomparison Project (CMIP), which can be used to train models to make global and regional climate projections. To access the NCEI's climate data, we used the Climate Data Online (CDO) portal to search and download data. The data can be programmatically accessed using the NCEI's Application Programming Interface (API).

Exploratory data analysis is a crucial step before utilizing large datasets for model training. The Dataset named NCDC storm events dataset is from United States Climate Data Center, NCDC. As a part of data cleaning, we removed missing and corrupted data, as well as checking for and resolving inconsistencies or errors in the data. For Data normalization, we scaled the data so that it has a mean of zero and a standard deviation of one, which can help to stabilize the training process for deep learning models and as a part of Data augmentation we created new data samples by applying various transformations to the existing data, such as rotating or shifting the data, which can help to increase the size and diversity of the training data set.

The dataset used in this work is a time series dataset from January 1950 to August 2019 providing critical information for flood observation. We have total of over 1 million samples in this dataset. We took 779,588 samples for training data which covers from 1950 January to 2012 December and 268,701 samples for testing data which covers 2013 January to 2019 August. Some of the features of the dataset are as follows: ["BEGIN_YEARMONTH", "BEGIN_DAY", "BEGIN_TIME", "END_YEARMONTH", "END_DAY", "END_TIME", "EPISODE_ID", "EVENT_ID", "STATE", "STATE_FIPS", "YEAR", "MONTH_NAME", "CZ_TYPE", "CZ_FIPS", "CZ_NAME", "WFO", "BEGIN_DATE_TIME", "CZ_TIMEZONE", "END_DATE_TIME", "INJURIES_DIRECT", "INJURIES_INDIRECT", "DEATHS_DIRECT", "DEATHS_INDIRECT", "DAMAGE_PROPERTY", "DAMAGE_CROPS", "SOURCE", "MAGNITUDE", "MAGNITUDE_TYPE", "FLOOD_CAUSE", "CATEGORY", "TOR_F_SCALE", "TOR_LENGTH", "TOR_WIDTH", "TOR_OTHER_WFO","TOR_OTHER_CZ_STATE", "TOR_OTHER_CZ_FIPS", "TOR_OTHER_CZ_NAME", "BEGIN_RANGE", "BEGIN_AZIMUTH", "BEGIN_LOCATION", "END_RANGE", "END_AZIMUTH", "END_LOCATION", "BEGIN_LAT", "BEGIN_LON", "END_LAT", "END_LON", "EPISODE_NARRATIVE", "EVENT_NARRATIVE", "DATA_SOURCE"]. After Data preprocessing and exploratory analysis, we identified 50 feature vectors to the proposed LSTM model. The proposed feature vectors are determined after finding the correlations between each set of feature vectors as shown in the fig. 1.

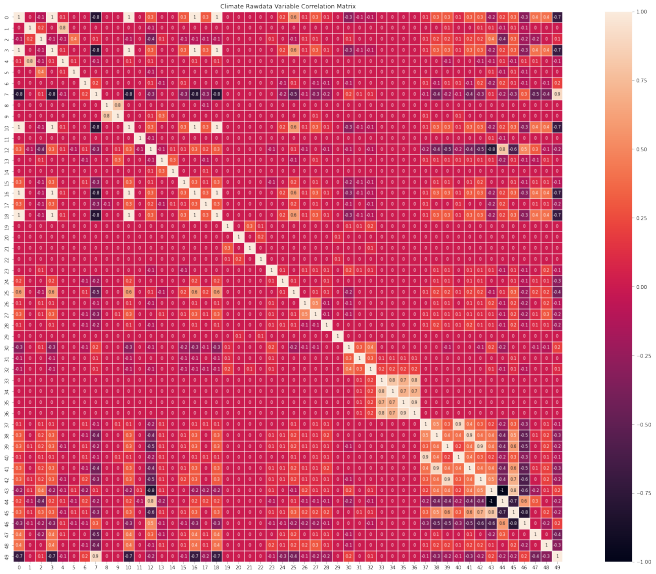

Figure 1: Correlation coefficient matrix of Feature Vectors

Methods

This paper investigates popular weather (time series-based) forecasting machine learning methods such as SVM, PCA and deep learning methods such as LSTM, for flood occurrence prediction. This work formulates the flood occurrence prediction as a classification problem.

Support Vector Machine (SVM)

Flood forecasting models have been achieving good results with Support vector machine (SVM) [12]. Typical architecture of SVM is shown in fig. 2. In this work, we chose radial basis function (RBF) kernel, and with grid search, we selected the parameters as listed in table 1.

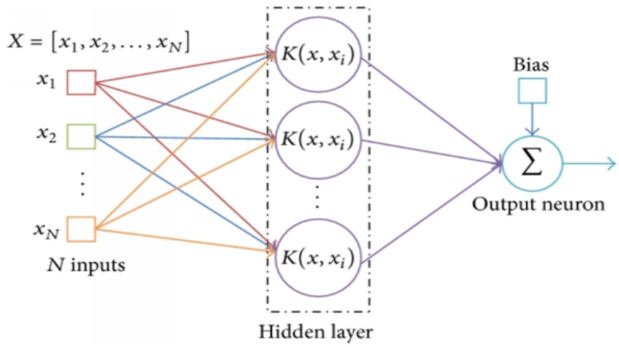

Figure 2: SVM model architecture

Long-short Term Memory-based Methods

Long Short-Term Memory network (LSTM) are a type of recurrent networks (RNN), that can store information over long periods of time in the dedicated memory cells. LSTM does not experience the exploding/ vanishing gradient problems that usual RNNs encounter. This allows them to learn long-term dependencies between input and output features. Due to the aforementioned characteristics, LSTMs perform exceptionally well in extreme-weather event problems where the timescale is typically long between the input and output.

Additionally, LSTMs have a demonstrated ability to model complex nonlinear feature interactions across numerous dimensions. Floods are extremely complex events and are caused by diverse factors that do not necessarily affect its water flow rate linearly. The said characteristics are thus crucial for accurate design of modern forecasting models. The model parameters for LSTM-based models are listed in table 1.

Principal Component Analysis (PCA): LSTM-PCA

The principal component analysis (PCA) [13] is a classical approach of feature extraction. Based on the data preprocessing and correlation coefficient analysis in data preprocessing section , PCA is applied to prune the indices of multiple correlating features (characteristics) to a few independent principal components. LSTM is further utilized for flood occurrence classification. LSTM-PCA architecture is represented in fig. 3.

Since there are high nonlinearity and hidden climate-related components in the NCDC data, the feedforward neural networks are unable to learn the complicate features of time series better than the neural networks with feedback connections. Thus, in order to strengthen the effectiveness of model, this work proposes a hybrid LSTM-PCA model to predict flood occurrence. The architecture in fig. 3 consists of three steps: (1) climate data preprocessing, (2) reducing dimensionality of influencing (highly correlated as in fig. 1) factors, (3) flood occurrence prediction by using the hybrid model. In step 1, the outliers are identified in the time-series and smoothed by using the weighted average method, respectively. For step 2, the PCA method is used to eliminate the unimportant features of influencing variables, as many of these variables are largely correlated with each other and thus leads to multicollinearity problems when training the model. The results of 1 and 2 steps determine the input variables of LSTM network in step 3.

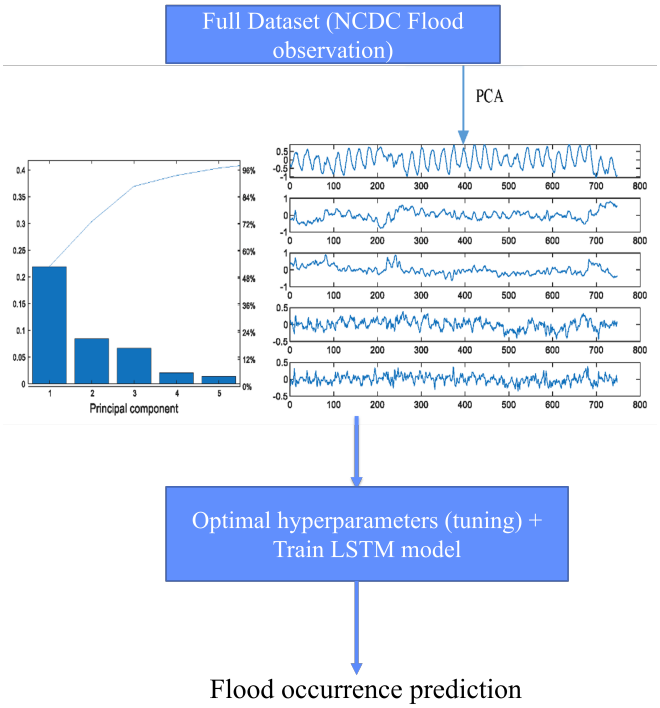

Figure 3: LSTM-PCA model architecture

Performance Metrics

To evaluate the performance of different models in this work, we utilized accuracy and root mean squared error (RMSE). Accuracy refers to the number of correctly classified (True positives (TP) and True Negatives (TN)) overall predictions made by the model, and is defined mathematically in equation 1:

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{1}$$

RMSE is calculated as the squared root of MSE, where MSE is the average value of the sum of squared differences between true yi and predicted $\hat{y}i$ values. Mathematically, RMSE is shown in equation 2:

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(yi - \hat{y}i)^2} \tag{2}$$

Evaluation Results and Discussion

In this section, the experimental results of the classifiers are discussed. Table 2 presents the classification results of the NCDC data set using the following classifiers: SVM, LSTM, and LSTM-PCA. The performance of each classifier is evaluated based on the ability of correctly classifying the time-series climate data in the test dataset for flood occurrence. The values for each model are calculated over 10 random samples generated by the 10-fold cross-validation.

The optimal structure of LSTM-based networks are selected by analyzing the hyperparameters sensitivities including the time steps of input variables, the number of hidden layers and the neurons in hidden layers. For the LSTM-PCA model, the inputs con-

Table 1: Model parameters.

Models	Hyperparameters
LSTM-based	Total training epochs: 100 LSTM Hidden state dimensions: 128 Number of cycles: 32 Learning rate: 0.001
SVM	Mostly set to defaults Penalty Coefficient: 1.0 Kernel Function Coefficient: 1/n where n is the number of features

Table 2: Average RMSE and accuracy comparison of models performance.

Models	Test Accuracy	RMSE
SVM	72.98	78.82
LSTM	92.54	74.96
LSTM-PCA	96.49	66.05

sist of the storms event and flood observation series after smoothing outliers using the weighted average method and the principal components of influencing factors. In the LSTM prediction model, the storm events + flood observations series without outliers and all the influencing features are inputted into model to train the unknown parameters of LSTM network.

The highest accuracy is obtained by the LSTM-PCA with the lowest RMSE. Figs. 4 and 5 present the precision-recall variations and model loss curve for the performant LSTM-PCA model.

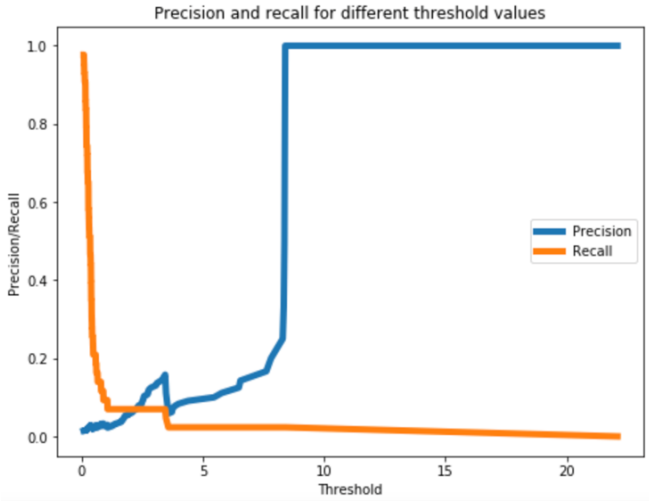

Figure 4: Precision and Recall Variations

Conclusion

The LSTM-based approaches achieve the highest classification performance of flood occurrence on the NCDC dataset with climate data and storm events. When comparing the performance of machine learning and deep learning models, LSTM-PCA outperforms LSTM by 4% imporvement in accuracy with decrease in RMSE by 12%. LSTM-PCA accuracy further improves by 32% and RMSE decreases by 16% as compared to SVM machine learning model. The models evaluated in this paper exhibit

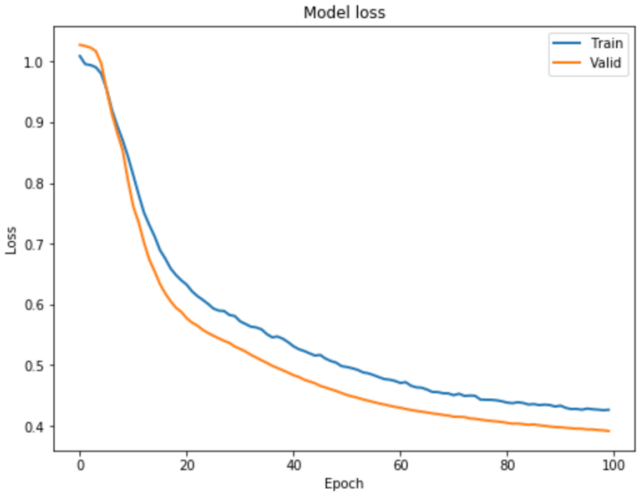

Figure 5: LSTM model loss

classification accuracies that are between 72%-97%. LSTM-PCA model has the highest performance.

In this work, we have investigated and compared different machine learning methods to analyze their performance in real and high- dimensional weather datasets. The experimental results show that the hybrid LSTM-PCA model combining with the PCA preprocessing technique produce rich input variables with stable variance and lower dimensionality.

References

[1] H. S. Munawar, A. W. A. Hammad, and S. T. Waller, "Remote sensing methods for flood prediction: A review," *Sensors*, vol. 22, no. 3, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/3/960

[2] M. N. Abdel-Mooty, W. El-Dakhakhni, and P. Coulibaly, "Data-driven community flood resilience prediction," *Water*, vol. 14, no. 13, 2022. [Online]. Available: https://www.mdpi.com/2073-4441/14/13/2120

[3] A. K. Lohani, N. Goel, and K. Bhatia, "Improving real time flood forecasting using fuzzy inference system," *Journal of Hydrology*, vol. 509, pp. 25–41, 2014.

[4] A. Mosavi, P. Ozturk, and K.-w. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, 2018. [Online]. Available: https://www.mdpi.com/2073-4441/10/11/1536

[5] Z. Han, J. Zhao, H. Leung, K. F. Ma, and W. Wang, "A review of deep learning models for time series prediction," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7833–7848, 2021.

[6] I. Gad and D. Hosahalli, "A comparative study of prediction and classification models on ncdc weather data," *International Journal of Computers and Applications*, vol. 44, no. 5, pp. 414–425, 2022. [Online]. Available: https://doi.org/10.1080/1206212X.2020.1766769

[7] S. Wang and J. Wang, "Research on prediction model of mountain flood level in small watershed based on deep learning," in *2022 4th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP)*, 2022, pp. 1024–1027.

[8] M. Moishin, R. C. Deo, R. Prasad, N. Raj, and S. Ab-

dulla, "Designing deep-based learning flood forecast model with convlstm hybrid algorithm," *IEEE Access*, vol. 9, pp. 50 982–50 993, 2021.

[9] P. Chaudhary, S. D'Aronco, J. Leitão, K. Schindler, and J. Wegner, "Water level prediction from social media images with a multi-task ranking approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 252–262, 2020.

[10] L. Yan, J. Feng, and T. Hang, "Small watershed stream-flow forecasting based on lstm," in *Proceedings of the 13th International Conference on Ubiquitous Information Management and Communication (IMCOM) 2019 13*. Springer, 2019, pp. 1006–1014.

[11] D. Niu, L. Diao, Z. Zang, H. Che, T. Zhang, and X. Chen, "A machine-learning approach combining wavelet packet denoising with catboost for weather forecasting," *Atmosphere*, vol. 12, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2073-4433/12/12/1618

[12] S. Li, K. Ma, Z. Jin, and Y. Zhu, "A new flood forecasting model based on svm and boosting learning algorithms," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, 2016, pp. 1343–1348.

[13] Y. Pang, D. Tao, Y. Yuan, and X. Li, "Binary two-dimensional pca," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 4, pp. 1176–1180, 2008.