Predicting Flood Severity in Indonesia based on Historical Flooding Events and Time-Series Information

Ankit Gupta[1], Abhir Karande[1], Rich Nguyen[1], Tho Nguyen[2]

[1] University of Virginia, Charlottesville, VA, USA
[2] National Academies of Sciences, Engineering, and Medicine
ag8mp@virginia.edu, ak9vqe@virginia.edu, nn4pj@virginia.edu, ThoNguyen@nas.edu

Abstract

Flooding is a key problem in Indonesia, but it's often difficult to find reliable data to serve as inputs to predict floods. In this paper, we provide a proof-of-concept automated system for predicting flood severity in Indonesia. We use four key indicators, which are tide level, precipitation, elevation and precipitation increase. The tide level and precipitation data are collected at hourly intervals, then are passed into an LSTM to predict tide and precipitation for future hours. The elevation and precipitation increase information are passed into a Dense Neural Network, which is trained on the percentage of people in a given area that are affected. The outputs of these two algorithms are two different weights, and depending on the predicted probabilities, a soft voting algorithm is used to assign a final flood severity score. This flood severity score, thus takes into account time-series based information, elevation, precipitation increase, and number of affected people. This research bridges a critical gap existing in environmental research applications, specifically by proposing a comprehensive, multi-level algorithm that has the potential to predict floods in Indonesia. This algorithm, when applied to the real world, has the potential to be deployed in Indonesia on a scalable and continuously collected dataset, to continuousy predict floods in Indonesia and allow for faster disaster response time in the future.

Introduction

Background

Indonesia is one of the countries most prone to flooding. According to a paper published by Nature, Indonesia is one of the top 5 most prone countries to flooding-based natural disasters [1]. Indonesia's location geographically, in having high levels of rainfall throughout the year, combined with extreme weather conditions, and the slope of the land being slanted at a larger angle, which results in potential flash floods throughout the area. Other sources of flooding include rivers in Indonesia overflooding, like the city of Birma in Indonesia. Another key problem is that 76 million people in Indonesia live on some land that can be considered to be high-risk for flooding.

Many countries that are affected by environmental factors like Floods do not have many resources, so helping predict flood severity for fast response can be extremely helpful. According to an article by Visual Capitalist, Indonesia's population is 27% at risk, and there are a total of 1.81 Billion people across the world that can be exposed to floods that would occur once in 100 years [2].

Motivation

As a result of these factors, one critical application of AI to advance science and engineering is to predict flooding in Indonesia. In order to do this, various different types of data needs to be used, each acquired from different datasets. This data needs to be aggregated, then different models need to be used to make sure that each type of data is being evaluated accurately in the prediction model. The purpose of this research paper is to present a proof-of-concept algorithm for flood prediction in Indonesia that is comprehensive, scalable, and uses a variety of types of data.

Related Work

One study, by Hapsari et al., focused on flood disaster management [3]. They found that the key solutions that could be used for flood management would be man-made factors, managerial issues, and flood disaster management. Another study, by Riza et al, was focused on mitigating flood disasters using Machine Learning [4]. This study focused on the potential of using machine learning to predict flooding based on rainfall, river water level, and river discharge. This paper mainly focused on conducting a literature review of past methods and relating them to how machine learning could be used, rather than specifying actual machine learning algorithms that could be used.

A third study, by Priyambodho et al, was focused on simulating extreme rainfall, using a rainfall model [5]. One key problem with this study is that they were not able to replicate past flood events and the system that they used, the GSMaP Gauge, was not able to be used to collect realtime data. This study used satellite rainfall data, but it only relied on rainfall, which is another limitation of this study.

A fourth study, by Miau et al, was focused on finding anomalies in flood data using deep learning [6]. Specifically, they looked at combining a CNN and GRU model to form a CNN-GRU model that can find patterns in forecasting river flooding levels in Taiwan. The main gap with this research is that it purely focused on river flood forecasting, rather than

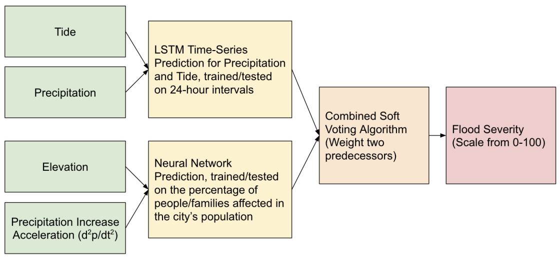

Figure 1: Different parts of the severity prediction machine learning model

.

focusing on a set of multiple factors that could contribute to flooding.

Clearly, no other research papers have focused on creating a comprehensive machine learning model to predict flooding in Indonesia. Research has generally focused on higher-level algorithms or simulating data, while more specific research focuses on one low-level feature rather than a set of features.

The goal of this research project is to bridge this gap by providing a proof-of-concept machine learning algorithm that considers multiple types of indicators of flooding in Indonesia, then uses different algorithms to create a comprehensive prediction for flood severity.

Methods

Overview

Figure 1, above, shows the general method used for predicting Floods in Indonesia. There are two different models, an LSTM and a DNN, that are used for predicting floods in Indonesia. Each model is trained on unique data, then a Soft Voting Method is used to combine the results to provide a final flood prediction score.

To begin with, we retrieved data from the United Nation's Office for the Coordination of Humanitarian Affairs, categorized data with severity, then created a dataset with flooding and non-flooding events. Next, we trained a classifier to predict flood severity given weather data, including precipitation and tide, for each of the events. By the end, we were able to achieve a MAE of 0.0043 based on a reliable dataset of flood events. This work can be integrated into our existing web application to predict flood severity categories given collected weather data. In the future, we plan to find hourly patterns associated with flooding, then classify whether or not floods occur. As a result, we will be able to predict flooding in Indonesia, then integrate these insights to help citizens across Asia be safer while increasing access to environmental flooding information across multiple countries.

Data Collection

The area of study for the flood forecasting in South-East Asia with a focus on the provinces throughout the country of Indonesia. For an effective time series prediction at multiple locations, there needs to be ample data at each of the locations, and in the context of flooding, ample data representing flood incidents in several locations around Indonesia.

Flood incidents affecting areas within Indonesia were compiled from articles from the United Nations Office for the Coordination of Human Affairs dating back to 2012. Due to inconsistencies in reporting style and provided information because of varying sources, the information that is collected from the articles is the effected locations and the date of impact. Upon compiling the various locations and dates of reports, the precipitation, elevation, and tidal data from those particular dates are accrued via the WeatherAPI and and StormGlass. For each report, the previous 24 hours of precipitation and tidal data were also recorded. The respective date, precipitation, elevation, and tidal data each serve as input variables for the forecasting. Additionally, for corresponding dates and locations, metrics including individuals affected were recorded. For each of the reports, the population of the corresponding location was found and the ratio of number of individuals affected to the population served as the severity of impact, the output variable.

Machine Learning

Machine Learning algorithms have become more prominent in recent years and are used for predictions and classifications for a wide array of applicable data sets. Supervised learning, a subset of machine learning, employs labeled training data to provide appropriate mappings of inputs to outputs in order to ultimately predict outcomes accurately. Supervised learning algorithms including a Long Short Term Memory Recurrent Neural Network and a Deep Neural Network will be trained on input data and used to provide a prediction for flood severity via a time series forecasting and prediction respectively.

For the Time-Series aspect, we propose using an LSTM that is trained on intervals of data from precipitation and tidal data. In order to train this model, for each day that there is a flooding event, we can retrieve the last 48 hours of data from that event, then set that data with an 80/20 split to train/test the model for tidal and precipitation prediction. Since higher tide and precipitation are directly correlated to flooding, these are two helpful indicators.

For the singular data point prediction, we propose using a Deep Neural Network, trained on elevation data and precipitation increases. We can use a normalization layer, dense layer, ReLu function, and an output dense layer. The output that can be used is the number of people affected at the population. As a proof of concept, we created an example neural network and trained it for 15 epochs.

Results

The DNN predicted flood severity for samples of the evaluation set with a mean absolute error: 0.004337307531386614. For context the mean of the outputs of the evaluation set is 0.0008258065857309899 and the standard deviation is 0.0016961530458533282. Provided below is an illustration of the error in predicting the flood severity:

Upon assessing the performance of the DNN model, the importance of the input variables (Rate of Increase in Precipitation and Elevation) in prediction was found. Permutation feature importance was performed on the training set,

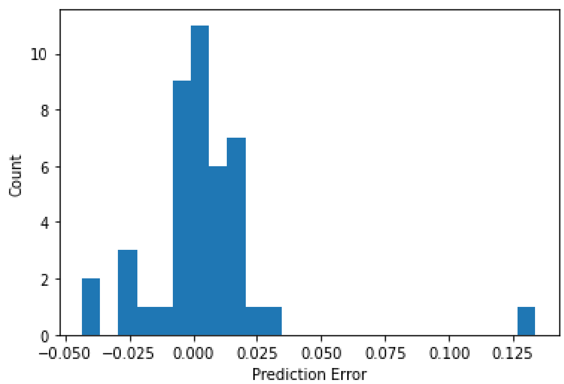

Figure 2: Prediction Error of DNN on Validation Set

.

in which the Elevation feature had a weight of 0.0121 and the Rate of Increase in Precipitation had a weight of 0.0079. From this the each of the can be described as having an importance to the model as performing permutations on the each feature's values changes the error of the prediction to a relatively more drastic extent as seen with the positive values.

Discussion

Compared to the related work, this paper bridges multiple gaps. To begin with, the paper provides the first-ever comprehensive machine learning algorithm for flood prediction in Indonesia. As mentioned in the related work, flood prediction is a critical problem that has been studied using techniques like looking at ways to mitigate floods, using machine learning algorithms at a high-level, and using simulation models. Some of the related work did focus on applying deep learning algorithms for prediction, but these studies were focused on just one indicator.

As opposed to multiple pieces of related work focused on the application of one algorithm to one piece of data, this work uses two different ML algorithms and four pieces of data. As opposed to using a singular data source that can only be used once, the data used for the proof-of-concept machine learning model can be retrieved from an API, meaning that data retrieval is easily scalable and can be applied to any other city in the world.

As opposed to purely focusing on time-series data, this machine learning algorithm considers taking historical data only from flood events, and also considers the percentage of the population that was affected from a given flood event. This focus on flood events rather than purely time-series based data gives this algorithm the ability to generalize to the severity of different types of flood events. This focus on collecting data from multiple flood events from the ASEAN database provides a unique angle on prediction that has rarely been used in past studies.

In other words, this research paper presents the first ever proof-of-concept method of using multiple distinct inputs, multiple different algorithms, and a soft voting algorithm to combine results to form a final flood prediction score.

Conclusion

Impact

This research presents the first-ever, comprehensive, multi-model, scalable method for flood prediction in Indonesia. The algorithm considers four different types of data, uses two different types of customized models, and uses a soft voting algorithm to combine the results of the models. The models are trained on data including the number of people affected, and the tidal and precipitation data can be collected hourly, to ensure that data is consistent. As opposed to relying on locally-sourced data, all of the data used for this proof-of-concept machine learning model can be collected online using the Weather API, the StormGlass API, and the ASEAN Natural Disaster Database.

The impact grows larger when we consider using this type of comprehensive flood prediction model for different levels of granularity and different countries. Given that this algorithm can work for any city provided, and the tide/elevation/-precipitation data can all be retrieved for any latitude/longitude, all of the data that was collected for this research study can also be collected for more comprehensive research studies in the future.

Future Work

There are several areas for future work. To begin with, the purpose of this paper was to provide an innovative method for predicting floods in Indonesia, but we still need to fully implement the method and achieve more results for each of the algorithms and the soft voting algorithm. Next, we can extend this algorithm to different countries beyond Indonesia, including other ones in Asia, to better understand whether the algorithm can work in other countries as well. Third, once the algorithm is fully implemented, we can look into adding more explainability towards the final flood prediction score, including revealing to the user the relative weighting of each of the individual outputs that are passed into the soft voting method.

Acknowledgements

This material is based upon work supported by the National Science Foundation Division Of Computer and Network Systems under Grant Award #2026050.

References

[1] Rentschler, J., Salhab, M., & Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. Nature communications, 13(1), 1-11.

[2] Floods in the neighborhood: Mapping poverty and flood risk in Indonesian cities. (2022, July 11). World Bank Blogs. https://blogs.worldbank.org/eastasiapacific/floods-neighborhood-mapping-poverty-and-flood-risk-indonesian-cities

[3] Hapsari, R. I., & Zenurianto, M. (2016). View of flood disaster management in Indonesia and the key solutions. American Journal of Engineering Research, 5(3), 140-151.

[4] Riza, H., Santoso, E. W., Tejakusuma, I. G., & Prawiradisastra, F. (2020, November). Advancing Flood

Disaster Mitigation in Indonesia Using Machine Learning Methods. In 2020 International Conference on ICT for Smart Society (ICISS) (pp. 1-4). IEEE.

[5] Priyambodoho, B. A., Kure, S., Yagi, R., & Januriyadi, N. F. (2021). Flood inundation simulations based on GSMaP satellite rainfall data in Jakarta, Indonesia. Progress in Earth and Planetary Science, 8(1), 1-17.

[6] Miau, S., & Hung, W. H. (2020). River flooding forecasting and anomaly detection based on deep learning. IEEE Access, 8, 198384-198402.