Forecasting sea level rise using machine learning techniques

Md. Riftabin Kabir, Nazmus Sakib Borson, Sifat Momen & Md. Sazzad Hossain*
Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
sazzad.hossain09@northsouth.edu

***Abstract*—During the past few decades, climate change has been posing as a vital game changer for the world stability of natural conditions. The effect can be easily demonstrated via the rise of sea levels on global and local scenarios. Increase of temperature, change in precipitation, melting of glaciers are causing the sea levels to rise in an alarming rate like never before. This particular paper focuses on predicting the sea level of Bangladesh, a third world South Asian regional country using advanced machine learning techniques to produce a potential model for future cautions. The proposed methodology uses climate data of previous 40 years (approx.) from 1977 to 2017 to train our model using different machine learning algorithms like Random Forest (RF), KNN and MLP. In testing phase, KNN algorithm prompted 91.3204% accuracy.**

***Index Terms*—Classification, Data Mining, Machine Learning, Sea Level, Prediction, Climate Change**

I. Introduction

Climate change is a natural phenomenon which may occur due to various natural processes. In the last 50-100 years climate change took a very drastic and dangerous turn [2] resulting in an array of many global problems. Global warming is one consequence of climate change. One of the major upsets is the melting of the glaciers which consequently results in the rise of sea level [7]. As almost half of the world's population lives near coastal region, this issue has become a serious matter of concern. It has therefore become imperative to be able to predict sea-level rise with high accuracy and precision. This paper uses machine learning approaches to develop models that can predict sea level rise with high accuracy.The lack of relevant climate data resources, has made the task of predicting sea level rise very challenging. This paper focuses on the sea level rise in the south Asian region, and in particular, the Bay of Bengal. To the best of our knowledge, there has been no work pertaining to the sea level rise in the Bay of Bengal.

II. Relevant Works

Most of the early studies concerning sea level prediction was based on dataset with yearly frequency distribution of climate data. Some researchers only used temperature data to reach the final conclusion. Very few work has been done regarding the Bay of Bengal. Some of the research work pertaining to Bay of Bengal include storm-surges prediction models [3] and discussion on the effects of sea-level rise in the coastal regions [9]. Unfortunately, none of them were precisely about the prediction of the sea-level rise of the Bay of Bengal. A paper on sea level prediction by Rahmstorf relates sea level to temperature which was published in Science journal [8]. Many researchers also use this semi-empirical approach to predict sea level rise. Their researches yield a wide range of variation on sea level rise ranging from 30 to 180 centimeters for the year 2100. Recently on an IPCC report,sea level rise was predicted using isostatic and tectonic effects correction [5]. A group from Stanford university worked on a project in their CS299 course,that predicts sea level rise using machine learning algorithms and error analysis methods [1]. They made the model for San Francisco Bay as well as global sea level.On a study by Yin land,ocean,sea-ice and atmosphere systems are integrated with climate model to predict sea level rise on the northeast coast of United States [10]. Some of the studies mentioned above have also introduced glacial isostatic adjustments to their models. Although there are noble studies lying in this domain, the authors propose some advancement and different approach in this paper to predict the sea-level rise of the Bay of Bengal, Bangladesh with high accuracy.

III. Methodologies

The methodology of our work is illustrated in figure 1.

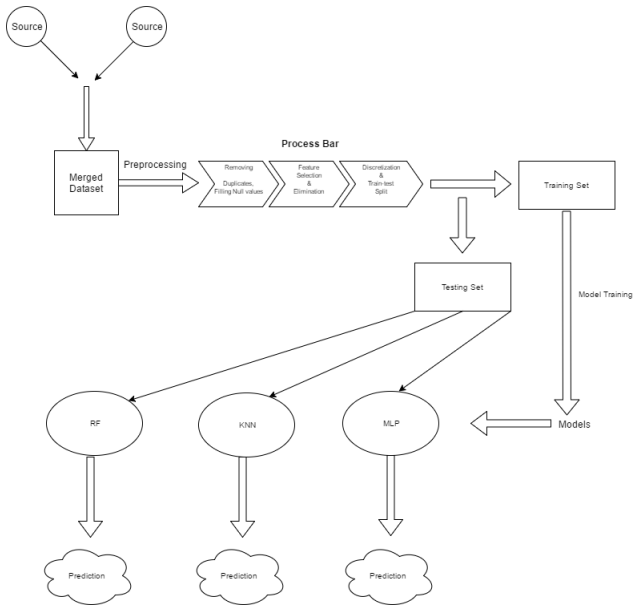

Fig. 1. Methods and Procedures

A. Data

The dataset is accumulated from two different sources and then were merged into one data-set. The sources are outlined in table I.

TABLE I
DATA-SET RESOURCES

Dataset	Resources
Climate	National Oceanic and Atmospheric Administration
Sea Level	University of Hawaii Sea Level Center

TABLE II
MERGED RAW DATA ATTRIBUTES AND DESCRIPTION

Variable Name	Description
YEARMODA	Year-Month-Date
LEVEL	Numerical Sea-Level
STN	Station Number
TEMP	Temperature
DEWP	Dew Point
SLP	Sea Level Pressure
STP	Station Pressure
VISIB	Visibility
WNDSP	Mean Wind Speed
MAXSPD	Maximum Sustained Wind Speed
GUST	Maximum Wind Gust
SNDP	Snow Depth
MAX	Maximum Temperature
MIN	Minimum Temperature
PRP	Precipitation
FRSHTT	Fog, Rain, Snow, Thuner and Tornado Indicator

The data-set contains 10828 instances of approximately 40 years from 1977 to 2017. It has 16 attributes including the class attribute which are as shown in the table II. After pre-processing the final attribute count was 10 including the class attribute as shown in table III.

TABLE III
FINAL FEATURE NAMES AND DETAILS

Feature Name	Description
YEAR	Year of the Event
MONTH	Month of the Event
CAT	Categorical Value of Level
TEMP	Temperature
WDSP	Wind Speed
SLP	Sea Level Pressure
DEWP	Dew Point
PRP	Precipitation
MAX	Max Temperature
MIN	Min Temperature

B. *Pre-processing*

As data is the core-component of machine learning algorithms, it is important to feed proper data to solve the problem accurately and more precisely. Without having properly groomed data it is not possible to obtain excellent or so to say desired results. For this, data pre processing is an absolute must. Preprocessing refers to the work flow of making the data clean, efficient and workable data. After the data is gathered from different sources it is initially in raw format which is not suitable enough for the required analysis. For this case preprocessing is an inevitable pathway for acquiring satisfying results. [6]

TABLE IV
REMOVED FEATURE NAMES AND DETAILS

Feature Name	Description	Reason
DATE	Date of the Event	Not significant
LEVEL	Sea Level	Converted to categorical
STN	Station Number	Not imapct
STP	Station Pressure	No impact
GUST	Maximum Wind Gust	No impact
SNDP	Snow Depth	No variation in instances
MAXSPD	Maximum Sustained Wind Speed	No Impact
FRSHTT	Fog,rain,snow,thunder,tornado indicator	No impact

1) Merging and Sorting : : After collecting the Sea Level data-sets, they were merged into a single sea level data-set. Then the merged sea level data-set was merged with the climate data-set. As the data-set was in two portions, it was necessary to merge them into a single data-set. To merge the two data-sets, 'YEARMODA' (Year Month Date) is used as primary key to match data exactly and compiled them into a single file. Then the data-set was sorted by date.

2) Removing Duplicates : : In the merged file there were redundant data due to multiple sea level data of same date from different weather stations. These duplicate instances had been removed using station id.

3) Filing Missing and Null Values by Mean/Median : : There were some missing data in the data-set. Those were filled up using mean value of the attributes.

4) Feature Selection and Elimination : : Some features in the data-set were not feasible enough to work with for the concerned purpose. For that reason the unnecessary features were removed which are presented in the table IV.

5) Discretization : : As classification models have been applied to predict the label, numerical sea level data had to be converted into categorical valuesV.The frequency distribution2 shows that the sea-level data were majorly divided into 4 groups ranging from 1200mm to nearly 4400mm. The groupings were done by 4 only to maintain the clarity and to avoid the over-simplification of the problem in discussion.

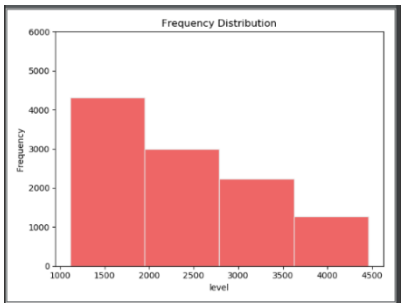

Fig. 2. Frequency Distribution of Sea-level Data

6) Train-test Split : : The Data-set was split into training and testing by the ratio 80:20 respectively. The training data contained 8662 instances and the testing set was consisted of the rest of the dataset. The split was done to ensure the unbiased performance regarding new and unseen data.

TABLE V
CATEGORICAL VALUE RANGE FOR SEA LEVEL

Sea Level Range	Categorical Value
0-1600	Low
1601 - 3000	Medium
3001 - 4000	High
above 4000	Very High

TABLE VI
CHANGED PARAMETERS OF RF

# of Iterations	Seed	Max-depth
1000	3	Unlimited

IV. ALGORITHMS

A unique approach was pulled out to predict the sea level rise in the upcoming years with categorical predictions rather than continuous values. For this, three very popular and advanced machine learning techniques were used which are Random Forest, KNN and Neural Network.

A. Random Forest

As an ensemble machine learning algorithm, this was one of our first choice to begin our classification. This algorithm uses varied sample sets to train each decision tree. It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the models prediction. The parameters of seed, number of iterations were varied to produce the optimum result. The parameters and fine tuning used are as of table VI.

B. KNN

Secondly, the K-nearest neighbors algorithm was used to train our model which uses the nearest neighbors to predict the class attribute. The K here is the number of neighbors to choose between.The number of K was changed and multiple experiments were done on with training set using this very effective classification algorithm. The changed parameters are shown in the table VII. The model used Euclidean distance as its distance matrix for which the formula is given below:-

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

C. Neural Network

Finally 'Neural Network' came into action to train our model. The multi-layer perceptron algorithm was implemented to classify the classes.A perceptron produces a single output based on several real-valued inputs by forming a linear combination using its input weights (and sometimes passing the output through a nonlinear activation function). Like this, many perceptron make up the MLP model.The model parameters that were varied were the seed number, learning rate(alpha) and the number of epochs to try out. The tweaking are parameters are as of table VIII. The equation of output function is as below:-

$$y = \phi(\sum_{i=1}^{n} w_i x_i + b) = \phi(w^T x + b)$$

TABLE VII
CHANGED PARAMETERS OF KNN

# of K	Distance Matrix
3	Euclidean

TABLE VIII
CHANGED PARAMETERS OF MLP

Learning Rate (Alpha)	# of Epochs	# of Hidden Layers
0.1	700	Arbitrary

V. RESULT ANALYSIS & EVALUATION

While building the model significant change of training time and accuracy were supervised as the parameters for different algorithms were being varied time to time to reach the optimum decision.

In RandomForest, during the training phase the model was gave out an accuracy of 93.0608% as training accuracy but in testing phase it increased and became a 93.8596% accurate model for the problem. The iteration number made an impact on the result as the model generated more trees to decide upon. The ROC curve 3 shows an almost very good behavior of the model trained. [4]

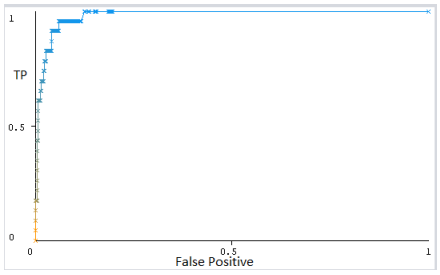

Fig. 3. ROC curve of the class "Very High" of RandomForest

While training the model with KNN, it was pretty straightforward to deal with. In testing period, the model rose up to 91.3204 % accuracy from just 90.9133% in training period. First the value of K was 1 which gave out a model with lower accuracy. But when the value of K was set to 7, significant increase in accuracy was demonstrated.

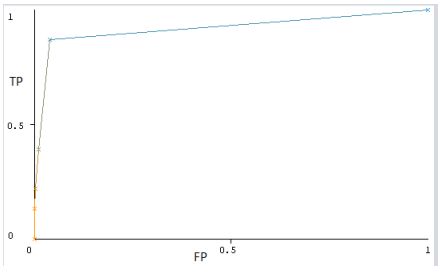

Fig. 4. ROC curve of the class "Very High" of KNN

Finally with the MultiLayerPerceptron, the accuracy level changed with a few parameters being tweaked. Like the other models, initially the training was done with only 500 epochs and a learning rate of 0.3 which made a model less accurate than the tweaked model. The learning rate, alpha then was lowered to 0.1 and the number of epochs were increased to 700. This little change made an increase in the accuracy of the model which was 91.3174% in training period and became 92.1976% in testing period. The slower learning rate and more epochs consolidated the training putting out a more accurate result than the initial one.

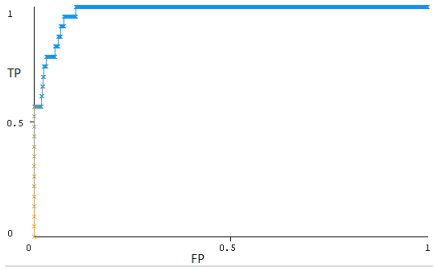

Fig. 5. ROC curve of the class "Very High" of MultiLayerPerceptron

However, after training the model with three different algorithms, the most accurate model was selected which was obviously the one trained with RandomForest algorithm because it presented the highest percentage of accuracy and it also has the lowest RMSE score which is the root mean squared error. Regardless to say, the other models also performed very well as expected. Although RandomForest resulted in the highest accurate model as we can see from table X, all the three models can be used to safely predict the future risks or so to say to forecast the change in sea levels of Bangladesh.

TABLE IX
DATA DIVISION FOR TRAINING AND TESTING

Data	Instances	Percentage
Training	8661	80%
Testing	2166	20%

TABLE X
MODEL VALIDATION

Model	Testing Accuracy	RMSE
RandomForest	93.8596 %	0.152
KNN	91.3204 %	0.177
Neural Network	92.1976 %	0.172

VI. CONCLUSION

Analyzing the results, it is impeccable to say that using random forest we were able to make the best prediction regarding sea-level rise.

Future enhancement of our research may include incorporating more salient features along with more diverse sources of data collection. Different algorithms and approach will be implemented to see the improvement of the model.

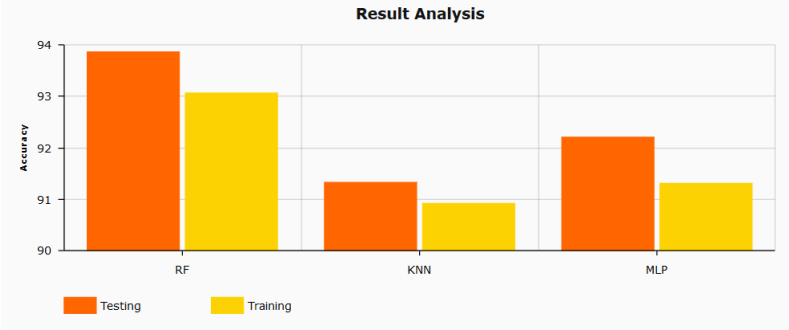

Fig. 6. Result Analysis and Comparison

REFERENCES

[1] Alahmadi, M., Kolmas, J.: Estimating the effect of climate change on global and local sea level rise (2015)
[2] Change, C.: Climate change. Synthesis report (2001)
[3] Das, P.: Prediction model for storm surges in the bay of bengal. Nature **239**(5369), 211 (1972)
[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1), 10–18 (2009)
[5] Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., González Rouco, J., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., et al.: Information from paleoclimate archives (2013)
[6] McKinney, W.: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference. pp. 51 – 56 (2010)
[7] Nicholls, R.J., Mimura, N.: Regional issues raised by sea-level rise and their policy implications. Climate research **11**(1), 5–18 (1998)
[8] Rahmstorf, S.: A semi-empirical approach to projecting future sea-level rise. Science **315**(5810), 368–370 (2007)
[9] Warrick, R.A., Barrow, E.M., Wigley, T.M.: Climate and sea level change: observations, projections and implications. Cambridge University Press (1993)
[10] Yin, J., Schlesinger, M.E., Stouffer, R.J.: Model projections of rapid sea-level rise on the northeast coast of the united states. Nature Geoscience **2**(4), 262 (2009)