Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm

Conference Paper · January 2021

DOI: 10.1109/ICREST51555.2021.9331199

3 authors, including:

Noushin Gauhar
Khulna University of Engineering and Technology
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Khadiza Sarwar Moury
Khulna University of Engineering and Technology
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm

Noushin Gauhar, Sunanda Das, Khadiza Sarwar Moury
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
noushingauhar.kuet.cse@gmail.com, sunanda@cse.kuet.ac.bd, khadiza.sarwar.moury@gmail.com

***Abstract*— Bangladesh is a flood-prone country. With limited resources and a major portion of the population living below the poverty line, flood impacts are severe. Deaths, malnutrition, widespread diseases, damage to infrastructure, disruption in the economy are some of the after-effects of this cataclysm. In order to put a flood management system into effect, it is essential to predict flooding events ahead of time. In this work, we applied different correlation coefficients for feature selection and k-nearest neighbors (k-NN) algorithm for the prediction of flood. The detailed result analysis shows that we achieved a high testing accuracy of 94.91%, average precision of 92.00% and an average recall of 91.00% using the k-NN machine learning model.**

Keywords*—Flood Prediction; Machine Learning; k-NN; Correlation Analysis; Data Scaling;*

I. INTRODUCTION

Due to the geographical location of the country in the Ganges Delta (also known as the Sundarbans Delta or the Bengal Delta), Bangladesh is highly susceptible to flooding. On average 20% of the country gets submerged causing loss of many lives, and economic loss [1]. The country has seen as worse as 66% of the land being submerged in an extreme flooding event [2].

Approximately 70.1% of the total land is used for agricultural purposes making Bangladesh an agriculture-based country [3]. Moderate flooding can submerge 20% of the country and extreme flooding may submerge 35% or more of the country [4]. Although moderate flooding brings alluvium to agricultural lands, extreme flooding results in major crop destruction. Moreover, a major decline in agricultural wages is seen as a result of the flooding event and around 1 million tons of crops get damaged because of floods [5, 6].

In Bangladesh, one in five people is living below the poverty line. Their homes get submerged and even get washed away by floods leaving them homeless. Many lose their jobs in flooding seasons. Flood causes many problems which include a scarcity of foods and clean drinking water. Access to sanitization and hygiene is compromised. Many waterborne diseases and contagious diseases increase with flooding events. Every year 5 million people are affected on average due to this calamity [7]. Public infrastructure damage negatively affects many people both directly and indirectly. Damage made to roads, bridges, vehicles, transportation systems can have profound effects on both the local and national economy. The communication system is abruptly disturbed. On average, the country sees financial loss worth more than 1 million dollars each year due to floods alone [6]. We cannot put an end to the aftermath of a flood altogether but with proper flood management, we can lessen the magnitude of it. For effective and efficient flood management, it is absolutely compulsory to be able to predict the flood first. In this work, we applied a machine learning algorithm i.e. k-nearest neighbors (k-NN) for this purpose. The features that are used in this work are of different ranges so a scaling function i.e. z-score normalization is employed to cope with the problem.

The rest of the paper is arranged as follows: Section II provides a brief background of some related research works. The methodology is illustrated in Section III. Section IV and Section V demonstrate the result analysis of the system and conclusions of the work respectively.

II. RELATED WORK

Many researchers have worked over the time to come up with effective methods to predict flood ahead of time to give Bangladesh enough lead time.

Chowdhury et al. [8] approached with finding the colinearity of Sea-surface temperature (SST) with a flood-affected area (FAA). They found the statistical relationship using principal component analysis. To construct a model to predict flood, multiple regression analysis was done. The statistical model uses SST, rainfall, and streamflow in Bangladesh to serve as predictors.

Shafizadeh-Moghadam et al. [9] have implemented eight different machine learning and statistical model and seven ensemble models for assessing flood susceptibility. Their dataset consists of 201 flood incidents in the Haraz watershed. They recognised eleven factors that contribute to flooding events. According to their evaluation parameters, the eight machine learning models: Artificial neural networks (ANN), Classification and regression trees (CART), Flexible discriminant analysis (FDA), Generalized linear model (GLM), Generalized additive model (GAM), Boosted regression trees (BRT), Multivariate adaptive regression splines (MARS) and Maximum entropy (MaxEnt) produce the Area under the ROC curve (AUC) scores of 0.920, 0.643, 0.822. 0.971, 0.962, 0.975, 0.941 and 0.971 respectively.

Han et al. [10] applied SVM models to forecast floods. They experimented with different kernel tricks and various input combinations. To assess the effectiveness of new models, they compared it with the Naive model, Trend model, and Transfer function (TF) model. For their study, they used the gamma values of 0.001, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.3, 0.5, 0.7 and 0.9. They applied 5-fold cross-validation for the training and testing phase. It was found that SVM predicted flood with higher accuracy in comparison to the TF model. Their observation showed that the linear function performs better with unknown future rainfall input. Extremely

978-0-7381-3042-2/21/$31.00 ©2021 IEEE

large rainfall data that is not scalable makes the model quite unstable.

Liong et al. [11] also applied SVM to develop models for flood stage forecasting. They used Artificial Neural Network (ANN) to compare their results. For their work, they used flood data of Dhaka, Bangladesh. The improvements they made in maximum predicted water level errors by SVM over ANN are given in cm: 9.6 (4 lead day), 22.6 (5 lead day), 4.9 (6 lead day) and 15.7 (7 lead day). They have pointed out advantages of SVM over ANN in details. They have used different combinations of C (regularization parameter) and σ value to get best performances by SVM. Tayfur et al. [12] attempted to predict flood hydrograph using machine learning methods such as ANN, genetic algorithm (GA), ant colony optimisation (ACO) and particle swarm optimisation (PSO). They used the data of Tiber River, central Italy, for their study. The data were used to train ANN as well as to find out the optimal parameter values of the rating curve method (RCM) for GA, ACO and PCO. The optimal value of α and β (RCM parameters) for GA are 1.22 and -5.86, for PSO are 1.20 and -5.90 and for ACO are 1.23 and -5.84. ANN, GA and ACO models made around 7 m^3/s MAE (Mean Squared Error) and 9 m^3/s RMSE (Root Mean Square Error) values on average. PSO model made 11 m^3/s MAE and 13 m^3/s RMSE on average. High errors were produced by RCM such as 13 m^3/s MAE and 17 m^3/s RMSE. They concluded that machine learning models required fewer substantial data and parameter estimations than physically based hydrologic models. Noymanee et al. [13] attempted to forecast pluvial floods using machine learning techniques. For this purpose, they used open data of the basin of Pattani River, Thailand. They used neural network (NN), bayesian linear (BL), boosted decision tree (BDT), decision forest (DF) and linear regression (LR). They split their dataset into two parts. They used the 80% of dataset for training the model and 20% dataset to test them. For the purpose of examining their models, they used RMSE, MAE and efficiency index (EI). They concluded that BL method provided the most efficient and effective result for a long-time lag. The other models also provided good results that were only acceptable for short time lag prediction.

Khosravi et al. [14] did a comparative study on decision trees algorithm to create flash flood susceptibility model. The mapping was done at the Haraz Watershed, Iran. Logistic model trees (LMT), reduced error pruning trees (REPT), Naive Bayes Trees (NBT) and Alternating Decision Trees (ADT) were used for the study. The historical data of 201 flash floods were used. For feature selection, they used information gain ratio and multicollinearity diagnostics method. 70% of the dataset was used for training the model and 30% of it was used to test the model. Statistical evaluation measures, the ROC curve and Freidman and Wilcoxon signed-rank tests were used to evaluate the performances of the four models. After analysing the results, it was seen that ADT model performed best.

III. Methodology

A. Dataset Preparation

The weather data of Bangladesh for 65 years is taken from Kaggle [15]. The data is from the Bangladesh Meteorological Department (BMD). The information on flood occurrence for a certain month and year was collected from a variety of sources including annual flood reports, newspapers, research papers, etc. and then merged with the weather data of BMD to create an updated dataset that can be found in [16] which consists of 20544 instances. The dataset contains information for 32 districts of Bangladesh. Some of the important attributes of the dataset include Rainfall, Cloud Coverage, Relative Humidity, Minimum Temperature, Wind Speed, etc.

B. Feature Selection

The main objective of this work is to predict flood. Feature selection is the procedure of selecting the appropriate features to be used in building of the model. Correlation analysis was performed for the feature selection process of the system. Correlation is the statistical relationship between any two variables. For feature selection, two types of correlation coefficient formulas were used:

a) Pearson correlation coefficient (r_{xy}): Measures the strength of existing linear relation between any two variables (e.g. x and y).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

Where n, x_i, $\bar{x}$, y_i and $\bar{y}$ respectively represents number of data points, value of x (i^{th} observation), mean of x, value of y (for i^{th} observation) and mean of y.

b) Spearman's rank correlation coefficient (r_s): Measures not only the strength but also the direction between any two ranked variables.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (2)$$

Where n and d_i represents number of cases and difference in paired ranks. After correlation analysis, the following equation was maintained in feature selection:

$$AND(r_{xy}, r_s) \geq 0.50 \quad (3)$$

Using Eq. (3), the following two features were selected from Table I: Rainfall and Cloud Coverage. The unrelated features may contribute to the model negatively giving unwanted results, thus they are not selected.

A. Data Scaling

Since the features of this dataset vary in units, range and magnitude, it was required to scale or normalize the data. In this work, we applied z-score normalization for this purpose. It is used to standardize the data by setting mean value to zero and scaling to unit variance.

$$z\_score = \frac{(x-\mu)}{\sigma} \quad (4)$$

where x, μ and σ respectively denotes sample, the mean of the training sample and the standard deviation of the training sample.

TABLE I. Correlation Analysis

Features	Pearson Correlation	Spearman rank Correlation
Rainfall	0.768816	0.662606
Cloud Coverage	0.573665	0.563592
Relative Humidity	0.459616	0.531344
Minimum Temperature	0.412115	0.437816
Wind Speed	0.217552	0.240350
Maximum Temperature	0.135808	0.115962
Bright Sunshine	-0.551552	-0.531379

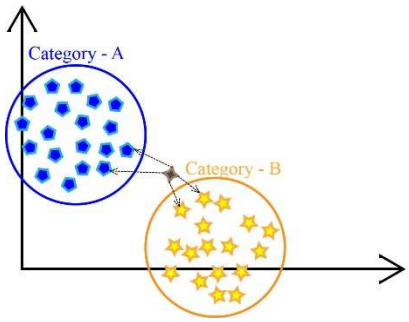


Fig. 1. Working principle of k-NN algorithm

B. ML Classifier :

k-Nearest Neighbor(k-NN): k-NN [17] is a popular supervised ML algorithm. It uses the property of feature similarity to make predictions of new datapoints. The predicted datapoints are assigned values depending on which points they match closest to in the training set.

The prediction of flood in Bangladesh using the k-NN algorithm follows as:

i. Determine the value of k, here k is the value of nearest neighbors
ii. Calculate the distance between training datapoints and the datapoint we want to classify
iii. Sort the training datapoints according to distance values in an ascending order
iv. Make prediction using the majority of nearest neighbors

We varied the value of k from 2 to 9 for this purpose. For assigning weights to the points, we used uniform weight function. This function weights all the points equally in every neighborhood. To calculate the distance, we used Minkowski distance formula. The Minkowski distance of order p between two variables X and Y can be defined as

$$\left(\sum_{i=1}^{n} |X_i - Y_i|^p\right)^{\frac{1}{p}} \quad (5)$$

Where n is number of the dimension of the variables. $p = 1$ is equivalent to the Manhattan distance and $p = 2$ is equivalent to the Euclidean distance. We used $p = 2$ in our system.

The Fig. 1 gives a visualization on how k-NN works to predict flood. The datapoint that needs to be classified is being compared to its k nearest points. It then evaluates which points are the closest and the most similar, and classifies the datapoint accordingly. In fig. 1, we see that the datapoint is closest and most similar to Category B (orange). So, the datapoint has been classified as Category B.

C. Training and Testing Phase

We split the entire dataset into a training and testing dataset with an 80:20 ratio. We set the test size = 0.2, which splits the 20% of the dataset into a testing set. The rest of the dataset remains as a training dataset. The random state was set to 50 to ensure that the train-test splits are always deterministic.

D. Evaluation Metrics

To evaluate the performance of the models, we used accuracy, precision, recall and f1-score. Here, the True Positive is a result where the model correctly predicts the actual flood class. Likewise, a True Negative is an outcome where the model accurately predicts the non-flood class. In False Positive, the model incorrectly predicts the flood class whereas in False Negative, the model incorrectly predicts the non-flood class. The following formulas are used to calculate the Accuracy, Precision, Recall, and F1-Score respectively.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative} \quad (6)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (8)$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

IV. Result Analysis

The system is implemented with python environment with the help of 'scikit-learn' library [18]. After feature selection and scaling, the training set is fed to the k-NN model. Table II. represents the performance measure indices of the system. From the Table II, it is evident that the highest accuracy, precision, recall and f1-score of the system are 94.91%, 92.50%, 91.00% and 92.00% respectively. For finding the best value of k, we plot the evaluation metrices for different values of k which can be found in fig. 2. According to fig. 2, the best result is achieved when k = 8. After k = 9, the value of accuracy and precision start decreasing.

TABLE II. Performance Measure Indices

	Accuracy	Precision	Recall	F1 score
k = 2	92.82	91.50	85.50	88.00
k = 3	93.41	89.50	90.00	89.50
k = 4	93.72	91.50	88.50	90.00
k = 5	94.28	91.00	91.00	91.00
k = 6	94.59	92.50	90.00	91.50
k = 7	94.60	91.50	91.00	91.50
k = 8	**94.91**	**92.50**	**91.00**	**92.00**
k = 9	94.79	92.00	91.00	92.00

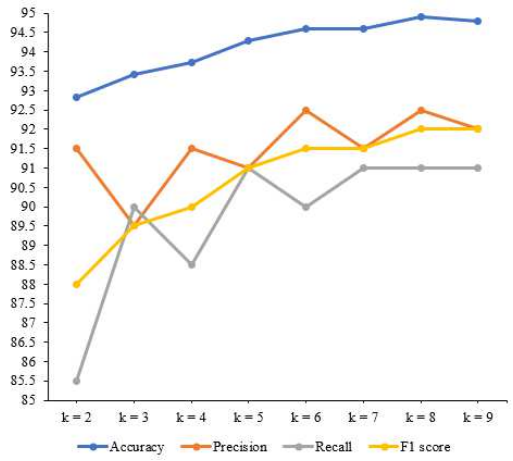
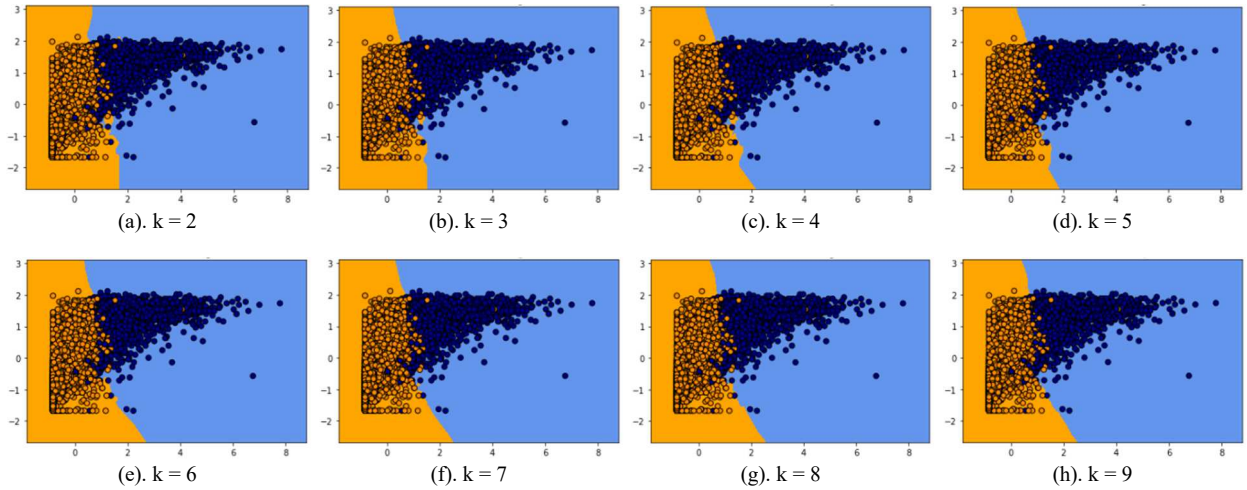


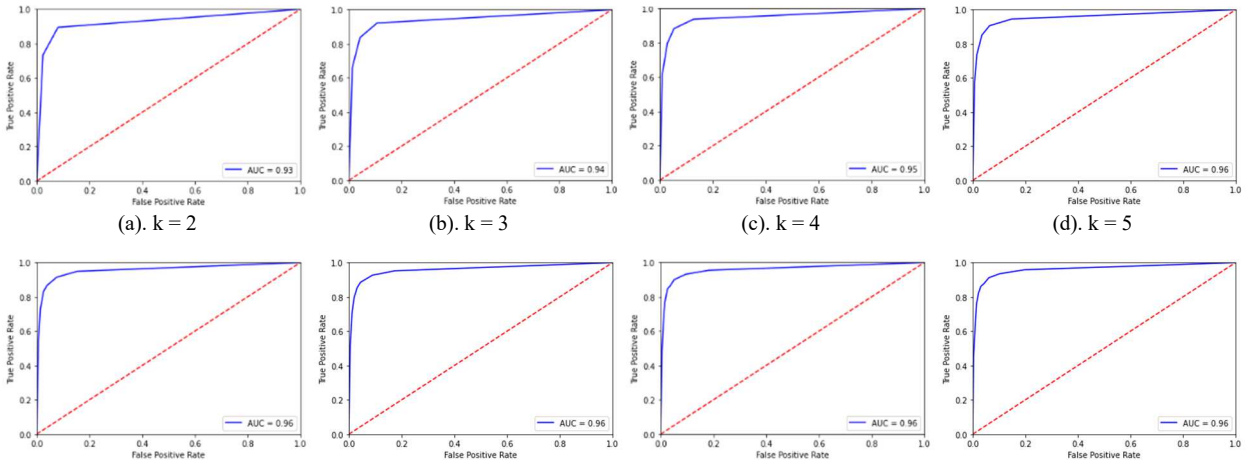
Fig. 2. Performance of the System for Different Values of k



(a). k = 2 (b). k = 3 (c). k = 4 (d). k = 5

(e). k = 6 (f). k = 7 (g). k = 8 (h). k = 9

Fig. 3. Decision Boundary Plot for Different Values of k where x and y axis represent Rainfall and Cloud Coverage respectively.



(a). k = 2 (b). k = 3 (c). k = 4 (d). k = 5

(e). k = 6 (g). k = 7 (g). k = 8 (h). k = 9

Fig. 4. Receiver Operating Characteristic (ROC) Curve for Different Values of k where AUC represents Area Under the Curve.

In fig. 3, we showed the decision boundary plot of the system for better understanding of the model. Fig. 4 represents the ROC curve of the testing phase for different values of k where, a larger area under the curve (AUC) is an indicator of better performance.

V. Conclusion

It is very important for Bangladesh to predict flood as accurately as possible because only then, Bangladesh will be able to alleviate the after-effects of floods. In this work, we applied correlation analysis, z-score normalization for feature selection and data scaling respectively. Finally, we used a k-NN model for the prediction purpose. The evaluation metrics of the system is calculated for different values of k so that we can determine the best value of k for the algorithms. The best value of k is 8 when the model produces optimum result with respect to accuracy, precision, recall and f1-score which are 94.91%, 92.50%, 91.00% and 92.00% respectively. In the future, we have a plan of applying other advanced machine learning algorithms for flood analysis and prediction.

The authors of this work believe that in the future, this work will contribute to predicting floods more accurately which can give Bangladesh the upper hand in the wake of a flooding event.

References

[1] A. M. Kamal, M. Shamsudduha, B. Ahmed, S. K. Hassan, M. S. Islam, I. Kelman, and M. Fordham, "Resilience to flash floods in wetland communities of northeastern bangladesh," *International journal of disaster risk reduction*, vol. 31, pp. 478–488, 2018.

[2] "After the flood: Official Damage Statistics of Bangladesh Flood 1998" reliefweb.com. https://reliefweb.int/report/bangladesh/after-flood-official-damage-statistics-bangladesh-flood-1998 (accessed Jan. 2, 2020)

[3] "Bangladesh Land Use." Indexmundi.com. https://www.indexmundi.com/bangladesh/land_use.html (accessed Jan. 2, 2020)

[4] L. Banerjee, "Effects of flood on agricultural productivity in bangladesh," *Oxford Development Studies*, vol. 38, no. 3, pp. 339–356, 2010.

[5] L. Banerjee, "Effect of flood on agricultural wages in bangladesh: An empirical analysis," *World development*, vol. 35, no. 11, pp. 1989–2009, 2007.

[6] D. AM, M. Nishigaki, and M. Komatsu, “Floods in bangladesh: A comparative hydrological investigation on two catastrophic events,” , vol. 8, no. 1,pp. 53–62, 2003.

[7] A. Dasgupta, “Floods and poverty traps: Evidence from bangladesh,” *Economic and Political Weekly*, pp. 3166–3171, 2007.

[8] M. R. Chowdhury and M. N. Ward, “Seasonal flooding in bangladesh–variability and predictability,” *Hydrological Processes: An International Journal,* vol. 21, no. 3, pp. 335–347, 2007.

[9] H. Shafizadeh-Moghadam, R. Valavi, H. Shahabi, K. Chapi, and A. Shirzadi, “Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping,” *Journal of environmental management,* vol. 217, pp. 1–11, 2018.

[10] D. Han, L. Chan, and N. Zhu, “Flood forecasting using support vector machines,” *Journal of hydroinformatics*, vol. 9, no. 4, pp. 267–276, 2007.

[11] S.-Y. Liong and C. Sivapragasam, “Flood stage forecasting with support vector machines 1,” *JAWRA Journal of the American Water Resources Association*, vol. 38, no. 1, pp. 173–186, 2002.

[12] G. Tayfur, V. P. Singh, T. Moramarco, and S. Barbetta, “Flood hydrograph prediction using machine learning methods,” *Water*, vol. 10, no. 8, p. 968, 2018.

[13] J. Noymanee, N. O. Nikitin, and A. V. Kalyuzhnaya, “Urban pluvial flood forecasting using open data with machine learning techniques in pattani basin,” *Procedia computer science*, vol. 119, pp. 288–297, 2017.

[14] K. Khosravi, B. T. Pham, K. Chapi, A. Shirzadi, H. Shahabi, I. Revhaug, I. Prakash, and D. T. Bui, “A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at haraz watershed, northern iran,” *Science of the Total Environment*, vol. 627, pp. 744–755, 2018.

[15] R. B. Reza. “65 Years of Weather Data Bangladesh Preprocessed” kaggle.com. https://www.kaggle.com/emonreza/65-years-of-weather-data-bangladesh-preprocessed (accessed Jan. 12, 2020)

[16] N. Gauhar. “Flood-prediction” github.com. https://github.com/n-gauhar/Flood-prediction (accesed Jun. 27, 2020)

[17] M.-L. Zhang and Z.-H. Zhou, “Ml-knn: A lazy learning approach to multi-label learning,” *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-sos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.