ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

A MACHINE LEARNING APPROACH FOR ESTIMATING RAINFALLUSING DIFFERENT DATA SOURCES

B. Naresh Kumar
Email: 0708.b@gmail.com
M. Tech, Department of Computer Science and Engineering.
Annamacharya Institute of Technology and Science, Hyderabad, Telangana, India.

Ramesh Babu Varugu
Email: rameshvarugu82@gmail.com
Assistant Professor
& HOD, Department of CSE.
Annamacharya Institute of Technology and Science, Hyderabad, Telangana, India.

Abstract - Providing an accurate rainfall estimate at individual points is a challenging problem in order to mitigate risks derived from severe rainfall events, such as floods and landslides. Dense networks of sensors, named rain gauges (RGs), are typically used to obtain direct measurements of precipitation intensityin these points. These measurements are usually interpolated by using spatial interpolation methods for estimating the precipitation field over the entire area of interest. However, these methods are computationally expensive, and to improve the estimation of the variable ofinterest in unknown points, it is necessary to integrate further information.

To overcome these issues, this work proposes a machine learning-based methodology that exploits a classifier based on ensemble methods for rainfall estimation and is able to integrate information from different remote sensing measurements. The proposed approach supplies an accurate estimate of the rainfall where RGs are not available, permits the integration of heterogeneous data sources exploiting both the high quantitative precision of RGs and the spatial pattern recognition ensured by radars and satellites, and is computationally less expensive than the interpolation methods.

Experimental results, conducted on real data concerning an Italian region, Calabria, show a significant improvement in comparison with Kriging with external drift (KED), a well-recognized method in the field of rainfall estimation, both in terms of the probability of detection. Accurate rainfall estimate is crucial for flood hazards protection, river basins management, erosion modeling, and other applications for hydrological impact modeling. To this aim, rain gauges (RGs) are used to obtain a direct measurement of intensity and duration of precipitations at individual sites.

Keywords – Computational infrastructure, geophysical data, GIS, oceans and water, radar data.

I. INTRODUCTION

Accurate rainfall estimate is crucial for flood hazards protection, river basins management, erosion modeling, and other applications for hydrological impact modeling. To this aim, rain gauges (RGs) are used to obtain a direct measurement of intensity and duration of precipitations at individual sites. In order to estimate rainfall events in areas not covered by RGs, interpolation methods computed on the basis of the values recorded by these RGs are used. Many variants of these methods have been proposed in the literature, and among them, the Kriging geo statistical method is one of the most used and recognized in the field.An accurate spatial reconstruction of the

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

rainfall field is a critical issue when dealingwith heavy convective meteorological events.

In particular, convective precipitations can produce highly localized heavy precipitation,not detected by sparse RGs, and floods can arise without a rainfall being detected. To overcome this issue, a recent trend in the literature is to integrate heterogeneous rainfall data sources to obtain a more accurate estimate by using interpolation Methods. A different approach relies on exploiting machine learning (ML) techniques. Typically, ensemble methods are used to address these issues. Ensemble is a classification technique, in which several models, first trained by using different classification algorithms or samples of data, are then combined to classify new unseen instances.

In comparison with the case of using a single classification model, the ensemble paradigm permits handling the problem of unbalanced classes and reducing the variance and the bias of the error. Especially, ensemble-based techniques can be used to address the issues concerning the rainfall estimation and to support the monitoring of meteorological events. These methods are also able to capture nonlinear correlations.

In order to address the main issues of rainfall estimation, in this article, an ML-based methodology, adopting a hierarchical probabilistic ensemble classifier (HPEC) for rainfall estimation, is introduced. The proposed approach, by integrating data coming from different source and exploiting an under-sampling technique for handling the unbalanced class's problem typical of this scenario, permits accurate estimation of the rainfall where RGs are not available.

Rainfall prediction using machine learning involves the use of historical weather data and other relevant factors such as temperature, humidity, wind speed, and pressure to train a model that can accurately predict future rainfall. The machine learning algorithms learn from patterns in the historical data to identify the relationships between these factors and rainfall..

II. EXISTING SYSTEM

An existing system is based on the ensemble paradigm include the work in which, similar to our work, employs a probabilistic ensemble and merges two sources of data even if the aim of this work is to develop a run-off analysis. Afterward, a blending technique is applied tothe results of the runoff hydrologic models to determine a single runoff hydrograph. Experimental results show that the hydrologic models are accurate and can help to make more effective decisions in the flood warning.

An evaluation of a real case study, located inthe European, proves the capability of the approach in providing accurate predictions for a hydrological partitioning of the region.

All these techniques are able to provide interesting results, but they require a rather delicate phase of parameters estimation of the particular model; therefore, as a side effect, usually, their flexibility and effectiveness tend to be hampered. As the relations between sensors data, cloud properties, and rainfall estimates are highly nonlinear, more flexibleapproaches based on ML techniques have been investigated recently. For instance, the problem of detecting convective events and closely related rainy areas is addressed in by using ANNs combined with support vector machines.

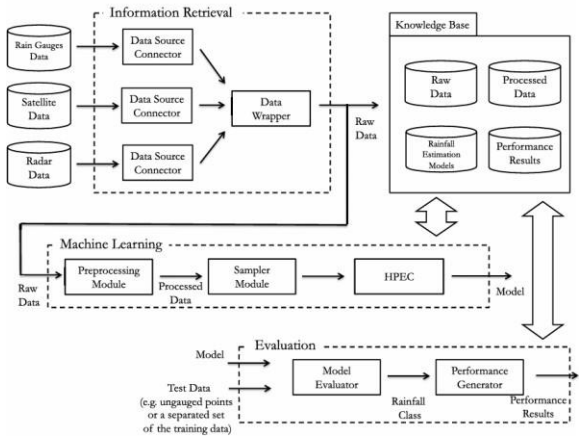
ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

III. PROPOSED SYSTEM

The proposed system is based on three heterogeneous data sources that is rain gauge, radar, and Meteosat are integrated to generate more accurate estimates of rainfall events. Different classification methods are compared on a real case data is taken from Calabria, a southern region in Italy, and a probabilistic ensemble approach is proposed.

Different ML-based methods, pre trained only on historical data, with a widely used interpolation method in the hydrological field are compared. Our approach is an effective solution for real scenarios, who has to analyze the rainfall in a specific zone presenting risks of landslides or floods. The experimental evaluation is conducted on real data concerning Calabria, a region located in the South of Italy, and provided by the DCP. Calabria is an effective test ground because of its strong climate variability and its complex orography.

IV. SYSTEM ARCHITECTURE:



V. IMPLEMANTATION

MODULES:

- Data collection
- Dataset
- Data preparation
- Predicting the rainfall in index calculation
- Model selection
- Analyze and prediction
- Accuracy on test set
- Saving the trained model

MODULES DESCSRIPTION:

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, them or and better data that we get, the better our model will perform. There are several techniques to collect the data, like web scraping, manual interventions. The dataset is located in the model folder.

Dataset:

The dataset consists of 1991individual data. There are 12 columns in the dataset, which are described below.

Date	01-12-20	03-11-21	23-08-20	24-10-21	06-04-20	11-06-21	16-04-20	09-09-21	13-11-20	02-07-21
Location	Hyderabad	Bangalore	Chennai	Mumbai	Cochi	Calcutta	Mangalore	Pune	Mysore	Delhi
MinTemp	13.4	13.1	14.1	15.7	16	NA	20.5	10.6	19	7.3
MaxTemp	22.9	35.4	29.2	29.7	29.5	14.8	23.5	15.4	24.9	12.7
Rainfall	0.6	0	0.2	0	1	0	3.2	6	0	0.6
Evaporation	NA	NA	2.2	7	NA	NA	0.2	1.6	8	1.8
Sunshine	NA	NA	7	11.1	NA	NA	0.2	NA	11.9	2.9
Temp	W	24.2	17.2	20	20	21.6	22.7	10.7	23.4	8.5
WindDir	20	W	NW	NNE	SE	ESE	NNE	NW	NE	WNW
Humidity	71	36	69	62	52	75	95	90	65	87
Pressure	1007.7	1010.5	1011.3	1018.6	1010.5	1017.3	1006.9	1009.8	1013.8	1011.2
Cloud	8	NA	6	5	7	NA	2	NA	5	7

OUTPUT DESIGN

Data preparation:

The first step is to collect the historical data, which includes the amount of rainfall and the corresponding values of the independent variables. Once the data has been collected, it needs to be cleaned and preprocessed to remove any outliers or missing values. Datapreparation for rainfall prediction typically involves collecting and processing various meteorological data sources. This can include

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

historical rainfall records, temperature, humidity, wind speed, atmospheric pressure.

VI. RESULT

Analyze and prediction: In rainfall prediction, ARIMA models can be used to analyses historical rainfall data and predict future rainfall based on trends and seasonal patterns. SVMs are machine learning models that can be applied to classification and regression applications

FIG 2 Accuracy Table

Algorithm	Accuracy
Linear Regression classifier	70.27%
Random Forest	97.85%
Naïve Bayes	71.21%
Support Vector Machine(SVM)	82.80%

Source	Name	Time res.	Space res.	Unit of meas.	Description
Radar	SRI	10 min.	1 km	mm/h	Surface rainfall intensity
	VMI	10 min.	1 km	dBZ	Maximum reflectivity on the vertical
	CAPPI2000	10 min.	1 km	dBZ	Reflectivity at the heights of 2000 m
	CAPPI3000	10 min.	1 km	dBZ	Reflectivity at the heights of 3000 m
	CAPPI5000	10 min.	1 km	dBZ	Reflectivity at the heights of 5000 m
MSG	Ch1	5 min.	4 km	K	$0.635\mu m$ channel brightness temperature
	Ch2	5 min.	4 km	K	$0.81\mu m$ channel brightness temperature
	Ch3	5 min.	4 km	K	$1.64\mu m$ channel brightness temperature
	Ch4	5 min.	4 km	K	$3.9\mu m$ channel brightness temperature
	Ch5	5 min.	4 km	K	$6.25\mu m$ channel brightness temperature
	Ch6	5 min.	4 km	K	$7.35\mu m$ channel brightness temperature
	Ch7	5 min.	4 km	K	$8.7\mu m$ channel brightness temperature
	Ch8	5 min.	4 km	K	$9.66\mu m$ channel brightness temperature
	Ch9	5 min.	4 km	K	$10.8\mu m$ channel brightness temperature
	Ch10	5 min.	4 km	K	$12\mu m$ channel brightness temperature
	Ch11	5 min.	4 km	K	$13.4\mu m$ channel brightness temperature
Rain gauge	x_i	-	punctual	m	x coordinate of the i^{th} rain gauge
	y_i	-	punctual	m	y coordinate of the i^{th} rain gauge
	$dist_i$	-	continuous	m	distance from the i^{th} rain gauge
	$rainfall_i$	1 min	punctual	mm	mm of precipitation

TABLE-I features extracted from the three sources of data: radar, satellite, and RGS.

Algorithm	CSI	FAR	POD	MSE
SVR	0.37 ± 0.010	0.40 ± 0.027	0.43 ± 0.011	0.11 ± 0.002
Decision Tree	0.41 ± 0.009	0.38 ± 0.033	0.47 ± 0.011	0.10 ± 0.002
Boosting	0.43 ± 0.008	0.33 ± 0.026	0.49 ± 0.008	0.09 ± 0.002
Random Forest	0.43 ± 0.010	0.31 ± 0.024	0.49 ± 0.011	0.09 ± 0.002
HPEC	0.44 ± 0.011	0.46 ± 0.016	0.58 ± 0.016	0.11 ± 0.002

Table II CSI, Far, Pod, and MSE for the SVR, Decision Tree, Boosting, Rf, And HPEC

Algorithm	Precision	Recall	F-measure
Class 4			
SVR	0.31 ± 0.04	0.07 ± 0.01	0.11 ± 0.02
Decision Tree	0.32 ± 0.08	0.08 ± 0.02	0.13 ± 0.03
Boosting	0.43 ± 0.07	0.09 ± 0.02	0.15 ± 0.03
Random Forest	0.47 ± 0.07	0.09 ± 0.01	0.15 ± 0.02
HPEC	0.24 ± 0.03	0.28 ± 0.03	0.26 ± 0.03
Class 5			
SVR	0.54 ± 0.14	0.11 ± 0.05	0.18 ± 0.08
Decision Tree	0.50 ± 0.14	0.18 ± 0.07	0.26 ± 0.08
Boosting	0.69 ± 0.11	0.24 ± 0.04	0.35 ± 0.05
Random Forest	0.68 ± 0.10	0.23 ± 0.05	0.35 ± 0.06
HPEC	0.33 ± 0.06	0.40 ± 0.07	0.36 ± 0.06

Table Iii Precision, Recall, and F-Measure for the SVR, Decision Tree, Boosting, RF, And HPEC for The Minority Classes 4 And 5

Algorithm	Precision	Recall	F-measure
Class 4			
Kriging	0.37 ± 0.042	0.20 ± 0.026	0.26 ± 0.031
Random Forest	$\mathbf{0.47 \pm 0.067}$	0.09 ± 0.014	0.15 ± 0.020
HPEC	0.24 ± 0.031	$\mathbf{0.28 \pm 0.030}$	0.26 ± 0.027
Class 5			
Kriging	0.45 ± 0.078	0.32 ± 0.059	0.38 ± 0.062
Random Forest	$\mathbf{0.68 \pm 0.101}$	0.23 ± 0.047	0.35 ± 0.057
HPEC	0.33 ± 0.062	$\mathbf{0.40 \pm 0.066}$	0.36 ± 0.057

TABLE V Precision, Recall, and F-Measure for the Kriging, Rf, And Hpec For The Minority Classes 4 And 5. The Values in Bold (Light Gray) Are Significantly Better (Worse) Than the Kriging Method

Algorithm	CSI	FAR	POD	MSE
All	0.44 ± 0.011	0.46 ± 0.016	0.58 ± 0.016	0.11 ± 0.002
No rain gauge	0.40 ± 0.005	0.53 ± 0.008	0.54 ± 0.010	0.16 ± 0.010
No radar	0.43 ± 0.007	0.47 ± 0.011	$\mathbf{0.58 \pm 0.012}$	0.12 ± 0.002
No meteosat	0.39 ± 0.006	0.52 ± 0.008	0.54 ± 0.012	0.16 ± 0.003

TABLE VI CSI, far, pod, and MSE for the HPEC, using all the features versus not considering, respectively, rg, radar, and satellite data. The values in light gray are significantly worse than the method using all the features

VII. CONCLUSION

An ML-based approach for the spatial rainfall field estimation has been defined. By integrating heterogeneous data sources, such as RGs, radars, and satellites, this methodology permits estimation of the rainfall, where RGs are not present, also exploiting the spatial pattern recognition ensured by radars and satellites. After a phase of preprocessing, a random uniform under sampling strategy is adopted, and finally, an HPEC permits the model used to be built to estimate the severity of the rainfall events. This ensemble is based ontwo levels: in the first level, a set of RF classifiers are trained, while, in the second level, a probabilistic metal earner is used to combine the estimated probabilities provided by the base classifiers according to a stacking schema. Experimental results conducted on real data provided by the Department of Civil Protection show significant improvements in comparison with Kriging with external drift, a largely used and well-recognized method in the field of rainfall estimation. In particular, the

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

ensemble method exhibits a better capacity in detecting the rainfall events. Indeed, both the POD (0.58) and the MSE (0.11) measures obtained by HPEC are significantly better than the values obtained by KED (0.48 and 0.15, respectively). As for the last two classes, representing intense rainfall events, the difference between the Kriging method and HPEC is not significant (in terms of F-measure) although HPEC is computationally more efficient.

VIII. FUTURE SCOPE

As future work, we plan to validate the method on a larger time interval, in order to consider effects due to seasonal and yearly variability, also considering the possibility of incrementally building the flexible ensemble model with the new data. In addition, we want to evaluate the effectiveness of the algorithm in individuate highly localized heavy precipitation events, also by adopting time series analysis to analyze the individual contributions of the different features for radar and Meteosat.

IX. REFERENCES

[1] J. E. Ball and K. C. Luk, "Modeling spatial variability of rainfall over acatchment," *J. Hydrologic Eng.*, vol. 3, no. 2, pp. 122–130, Apr. 1998.

[2] S. Ly, C. Charles, and A. Degré, "Different methods for spatial interpolationof rainfall data for operational hydrology and hydrological modelingat watershed scale. a review," *Biotechnologie, Agronomie, Société etEnvironnement*, vol. 17, no. 2, p. 392, 2013.[3] H. S. Wheater *et al.*, "Spatial-temporalrainfall fields: Modelling andstatisticalaspects," *Hydrol. Earth Syst. Sci.*, vol. 4, no. 4, pp. 581–601,Dec. 2000.

[4] J. L. McKee and A. D. Binns, "A review of gauge–radar mergingmethods for quantitative precipitation estimation in hydrology," *CanWater Resour. J./Revue Canadienne des RessourcesHydriques*, vol. 41,nos. 1–2, pp. 186–203, 2016.

[5] F. Cecinati, O. Wani, and M. A. Rico-Ramirez, "Comparing approachesto deal with non-gaussianity of rainfall data in Kriging-based radargaugerainfall merging," *Water Resour. Res.*, vol. 53, no. 11,pp. 8999–9018, Nov. 2017.

[6] H. Wackernagel, *Multivariate Geostatistics: An Introduction With Applications*Berlin, Germany: Springer, 2003.

[7] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2,pp. 123–140, Aug. 1996.

[8] B. J. E. Schroeter, *Artificial Neural Networks in Precipitation Now-Casting: An Australian Case Study*. Cham, Switzerland: Springer, 2016,pp. 325–339.

[9] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "ConvolutionalLSTM network: A machine learning approach for precipitationnowcasting," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1,Dec. 2015, pp. 802–810.

[10] W.-C. Hong, "Rainfall forecasting by technological machine learningmodels," *Appl. Math. Comput.*, vol. 200, no. 1, pp. 41–57, Jun. 2008.

[11] A. Parmar, K. Mistree, and M. Sompura, "Machine learning techniquesfor rainfall prediction: A review," in *Proc. 4th Int. Conf. Innov. Inf.EmbeddedCommun. Syst. (ICIIECS)*, Mar. 2017, pp. 152–162.

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

[12] M. Lee, N. Kang, H. Joo, H. Kim, S. Kim, and J. Lee, "Hydrologicalmodeling approach using radar-rainfall ensemble and multi-runoff-modelblending technique," *Water*, vol. 11, no.4, pp. 1–18, 2019

[13] C. Frei and F. A. Isotta, "Ensemble spatial precipitation analysis fromrain gauge data: Methodology and application in the European alps,"*J.Geophys. Res., Atmos.*, vol. 124, no. 11, pp. 5757–5778, Jun. 2019.

[14] J. L. Peña-Arancibia, A. I. J. M. van Dijk, L. J. Renzullo, and. Mulligan, "Evaluation of precipitation estimation accuracy inreanalyzes, satellite products, and an ensemble method for regions inAustralia and south and East Asia," *J. Hydrometeorology*, vol. 14, no. 4,pp. 1323–1333, Aug. 2013.

[15] V. Levizzani, P. Bauer, and F. J. Turk, *Measuring precipitation fromSpace: EURAINSAT and the Future*, vol. 28. Berlin, Germany: Springer,2007.