Binary Rainfall Classification using SMOTE: An Effective Machine Learning Strategy

Syed Atif Moqurrab[1], Abdul Razaque[1], Yersain Chinibayev[1] and Tolganay Chinibayeva[1]

[1]*International Information Technology University, Manas St. 34/1, Almaty, 050040, Kazakhstan*

Abstract

Water is critical to human survival. The plants we grow, the animals we boost, and the requirement human body to stay hydrated all depend on water. It is crucial to precisely predict the rainfall for effective utilization of water resources, food productivity, and proper storage of water. Because of recent climate changes, accurate rainfall forecasting has become more complicated than earlier. This paper improves the efficiency and accuracy of rainfall forecasting with the help of data balancing through SMOTE and machine learning. The dataset of twelve years duration was collected from a weather forecasting portal which includes several atmospheric attributes. Preprocessing methodologies are applied first, which include cleaning and normalization of data as well as data balancing using SMOTE. Performance comparison has been made for various machine learning techniques which include Naïve Bayes, MLP SVM, KNN, and Decision. It has been found that Decision Tree outperforms other techniques in terms of forecasting accuracy, precision, recall, and f1 measures. The best accuracy we achieved using the Decision Tree in this research was 99.8% for both rain and no rain classes. Similarly 100% precision,99% recall, and 99% f1measure for no rain class and for rain.

Keywords

Forecasting, SMOTE, Data Cleaning, Normalization, Data Balancing, Framework

1. Introduction

Rain forecasting has applications in water storage and management, flood prevention, agricultural planning, mobility planning, and many other fields. Accurate rain forecasting is an important and complex research area. Supervised machine learning techniques have been mostly used in the literature for this problem. There are various environmental factors such as humidity, wind speed, pressure, concentrations, and pollutants that have a role in rainfall. In the past, many researchers have investigated and proposed various methodologies and algorithms for predicting rainfall and are still engaged in this research area for improved results in terms of efficiency and accuracy using data mining and machine learning techniques. Machine learning algorithms make use of time series data by analyzing it for rain prediction.

Time series analysis is an approach for the creation of accurate models with the values of the variables positioned at periodic intervals [1]. Reading of time series data supports understanding of unseen forms of the data and assists in improved examination by using a suitable model for effective prediction. Time series data is normally gathered over a certain time duration on regular intervals [2-5] and can be used for forecasting in multi-domain areas like economic conditions, stock exchange, and weather, etc. However, weather forecasting with the help of time series data is a complex job [6-8].

Another method for rainfall forecasting is via statistical methodology, however, it requires lots of data attributes like local time, seasons, air pressure, cloud conditions, temperature, humidity, etc. As the nature of rainfall data is non-linear which makes the data noisy and unbalanced, various

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

a.razaque@iitu.edu.kz (A. Razaque); y.chinibayev@iitu.edu.kz (Y. Chinibayev); t.chinibayeva@iitu.edu.kz (T. Chinibayeva)

0000-0003-3284-1755 (S.A. Moqurrab); 0000-0003-0409-3526 (A. Razaque); 0009-0009-8985-5892 (Y. Chinibayev); 0000-0002-2657-3697 (T. Chinibayeva)

techniques need to be applied like data cleaning, normalization, and balancing on it to achieve higher accuracy in results.

Weather forecasting can help to take necessary measures to avoid human, animal, and infrastructure losses and to support the development of agriculture, economy, and health of any country and its people. In this research, we have performed data processing using SMOTE (Synthetic Over-sampling Technique Minority) and applied various machine learning algorithms to achieve higher accuracy and efficiency in results as compared to previously performed experiments by various researchers using data mining techniques [9-12]. For our experiments, we have used the rainfall data in Lahore – City of Pakistan for over 12 years (December 2005 to November 2017) [13].

For predicting rainfall, a classification framework is applied where the datasets are initially processed through cleaning, normalization, and balancing. It has been observed that datasets normally contain inaccurate or omitted values. Using data cleaning, such anomalies can be removed. Unclean data can lead to a range of issues, including linking errors, model misspecification, errors in parameter assessment, and wrong examination that in return results in false conclusions. Whereas, normalization is a process that is frequently used to prepare data for machine learning. The objective of normalization is to transform the values of numeric columns into a common scale in the dataset to refer to it, without distorting range differences or losing information. These pre-processing steps are essential for a smooth classification method with a high rate of accuracy [14, 15]. It has been found that Decision Tree outperforms other techniques in terms of forecasting accuracy, precision, recall, and f1 measures. The best accuracy we achieved using the Decision Tree in this research was 99.8% for both rain and no rain classes. Similarly 100% precision,99% recall, and 99% f1measure for no rain class and rain.

The organization of the paper is as follows: Section II discusses related work in the field of rainfall prediction using machine learning, section III describes the proposed methodology and techniques adapted, section IV defines the dataset used and its pre-processing, followed by the experiments performed and results tabulated in Section V. Section VI concludes the paper.

2. Related work

There are numerous methodologies and algorithms for data mining and machine learning to forecast rainfall [9-15]. However, we have investigated only those that are closely related to our approach. Some researchers [16,17] have used a neural network model by capturing non-linear dependencies of past weather modes and future climate states. Some other researchers have used support vector machines [18] to classify directly for environmental forecasting, however using SVM, the results get limited in range as compared to neural network methodologies. Few other techniques have also used Bayesian networks to model and forecast weather [19]. The technique adapted uses a machine-learning algorithm to find the most prime Bayesian networks and factors to reduce the computation cost based on different dependencies and the experiments have shown promising results. In general, in the domain of forecasting and visualization of huge collections of datasets, SOM (Self-Organizing Map) and Support Vector Machine are the prime machine learning methodologies.

Generally, the experiments for weather forecasting use a two-step approach. First, the dataset is split into a tiny set of vectors, then these vectors are divided into teams using victimization clump algorithms. The main objective of hierarchical algorithms is to scale back process prices for every cluster. The second step provides a rough image for every cluster thus lowering the prediction, and cost and increasing dependability [20] as compared to other techniques. The researchers have used a similar approach in their experiments using a number of different machine learning algorithms. Comparative analysis has been performed for various machine learning techniques such as M5 Model Trees, Support Vector Machine, Logistic Regression, Markov Chain, Radial Basis Neural Network, Genetic Programming, and k-Nearest Neighbor [21] for rainfall forecasting time series data of 42 towns using numerous climate attributes. The research verified that machine learning algorithms can perform well compared to the Markov Chain

methodology. There are several other models recommended by the researchers [22-28], however accurate forecasting examination has not been achieved because of the difficult data structures of the weather, categorical and dynamic patterns of the weather, noise in data, and dimensionality of the data. Therefore, there is a requirement for an effective model to predict the weather.

The latest study [9] for rain forecast prediction did a comprehensive analysis of binary classification (Rain and No rain). However, the author uses multiple well-known classifiers with different data splitting ratios for both the 'No rain' and 'Rain' classes. Based on their result the 'No rain' class predicted with high precision, recall, and F1 measure respectively as compared to the 'Rain' class. To improve the rain class prediction was their future research work. In this paper, we limit our research to improving the rain class prediction. The details of our proposed methodology are available in section 3.

Next, we have proposed our methodology that provides higher accuracy in rainfall prediction.

3. Methodology

The objective of this research is to compare, validate, verify, and receive higher accuracy in the result for forecasting rainfall in Lahore - a City in Pakistan using effective techniques, such as SMOTE and Machine learning.

The methodology adopts a three-step approach. The first step provides pre-processing on selected datasets by applying data cleaning, data normalization, and data balancing using SMOTE. The second step is applying machine learning algorithms to train and classify data. The algorithms used in our experiments are Naïve Bayes, Logistic Regression, SVM, KNN, MLP, Decision Tree, and Random Forest. Step three tabulates and evaluates results using accuracy, precession, recall, F1 measure, TP rate, and FP rate. The complete methodology is shown in Figure 1.

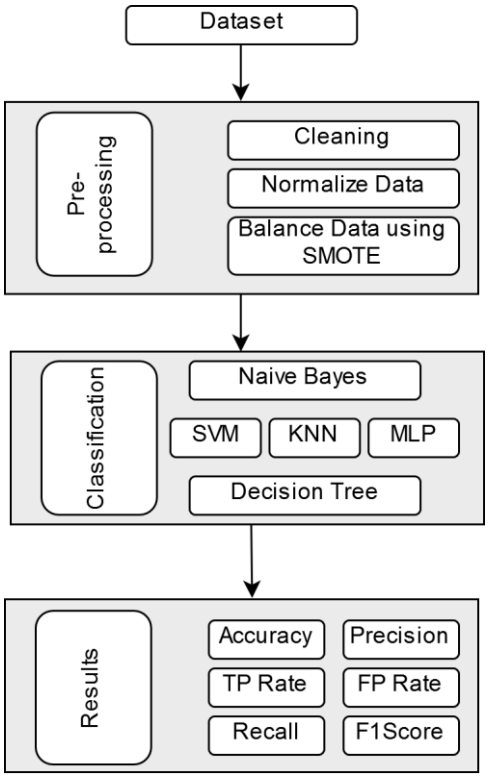

Figure 1: Proposed Methodology

4. Dataset and pre-processing

Time series models are the premise for any study of the performance of procedures over certain time period. Time series prediction is a significant region of machine learning [29]. In our

methodology, we have used datasets developed using time series models. The dataset includes many environmental attributes. Table I describes various attribute used, their types and their units of measurement.

Table 1
Features in Dataset

No	Attributes		
	Name	Type	Unit of Measurement
1	Temperature	Continuous	Degree Celsius
2	Atmospheric Pressure (Weather Station)	Continuous	Millimeter of Mercury
3	Atmospheric Pressure(Sea Level)	Continuous	Millimeter of Mercury
4	Pressure Tendency	Continuous	Millimeter of Mercury
5	Relative Humidity	Continuous	%
6	Mean Wind Speech	Continuous	Millimeter of Mercury
7	Minimum Temperature	Continuous	Degree Celsius
8	Maximum Temperature	Continuous	Degree Celsius
9	Visibility	Continuous	Km
10	Dew Point Temperature	Continuous	Degree Celsius

Typically dataset contains misleading values. Data cleaning process helps in recovering the missing values. Missing value can create inaccuracy in results. In the cleaning process, we have replaced the missing values with the mean which is one the most widely used methods in the literature. Table II shows Valid and Missing records in each attribute with the selected dataset.

Table 2
Attributes with Valid and Missing Values

No	Attributes		
	Name	Valid Records	Missing Values
1	Temperature	25,846	73
2	Atmospheric Pressure (Weather Station)	23,689	2,230
3	Atmospheric Pressure(Sea Level)	23,714	2,205
4	Pressure Tendency	11,320	14,599
5	Relative Humidity	25,790	129
6	Mean Wind Speech	25,890	29
7	Minimum Temperature	2,415	23,504
8	Maximum Temperature	4,174	21,745
9	Visibility	25,829	90
10	Dew Point Temperature	25,865	54

Missing values were replaced by mean, and data normalization was applied to maintain the values in certain boundaries [1, 11]. This normalization is performed using Z-Score: a commonly used mythology for this purpose. The normalization approach deals with the noise via prescribing the values intervals. However, after this missing value replacement and normalization, the dataset still contains discrepancies i.e. the data is highly imbalanced. This imbalance means that one class is represented using a large number of instances whilst the other is represented by a handful instance [30]. Thus, the data is required to be balanced.

There are many techniques available to balance the distribution of the classification type variable. In our experiments, we have used SMOTE (Synthetic Minority Oversampling method) because of its extensive use for data balancing in the literature [30]. SMOTE is a technique that reduces the effect of getting a few times inside the minority elegance. The strategy includes taking a subset of records from the minority elegance, intelligently growing new synthetic comparable

times, adding them to the authentic dataset, and using the brand new dataset as a sample in the schooling procedure for the classifier version [31].

4.1 Classifiers

Various classifiers were used in this research which is discussed in details in section 4.1.

4.1.1 Naïve Bayes

Naive Bayes (NB) classifier expect that the nearness of a specific component in a class is inconsequential to the nearness of some other element. Equation 1 and 2 shows the working of Bayesian classifier.

$$\boldsymbol{P(f/Z) = \frac{P(Z/f)P(f)}{P(z)}} \quad (1)$$

$$\boldsymbol{P(f/Z) = P(Z1|f) \times P(Z2|f) \times \cdots \times P(Zn|f) \times P(f)} \quad (2)$$

$P(f/Z)$ Represents the probability of class (f) given the predicator (Z);
$P(f)$ Shows the probability of class;
$P(Z)$ Shows the probability of Predicator;
$P(Z/f)$ Represents Likelihood ratio of predicator class.

4.1.2 KNN Classifier (KNN)

To predict new data points, the K-Nearest Neighbor Classifier (KNN) uses a similarity measures approach. The reason this study uses the KNN algorithm is that it depends entirely on the resemblance of the characteristics. Selecting the correct value of K is very essential to obtain ideal outcomes. K's value is the amount of closest neighbors regarded in a vector's classification.

$$\textbf{Eculidean Equation} = \sqrt{\sum_{i=1}^{n}(Yi - Zi)^2} \quad (3)$$

$$\textbf{Manhattan Equation} = \sum_{i=1}^{n}|Yi - Zi| \quad (4)$$

$$\textbf{Minkowski} = \left(\sum_{i=1}^{n}(|Yi - Zi|)^q\right)^{1/q} \quad (5)$$

Above mentions equations represents the similarity level between two data points'. Yi and Zi represent "n" data points.

4.1.3 Support Vector Machine (SVM)

Support Vector Machine "(SVM) is a supervised algorithm for machine learning that can be used for classification or regression challenges. It is mostly used in classification issues, though. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of characteristics you have) with the value of each function being the value of a specific coordinate. Then, by discovering the hyper-plane that differentiates the two classes very well.

$$\frac{1}{2}T^w T + \alpha \sum \varepsilon_i \quad (6)$$

$$X_i\,(T^w \emptyset\,(y_i) + c) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \geq 0, i = 1, \ldots\ldots n \quad (7)$$

Where α is a steady capacity, T is a coefficient vector, c is a constant and ε_i represents parameters for the handling of non-input information. The index I marks the instances of N practice. Note that the class labels are represented by $X \in \pm 1$ and the independent variables are represented by y_i. The kernel is used to convert information into the function space from the

input (independent). It should be observed that the greater the α, the greater the penalization of the mistake. α should, therefore, be carefully selected to prevent overfitting.

4.1.4 Decision Tree

The Decision Tree (DT), another algorithm used in latest anomaly-based IDS studies, is the same as any tree structure composed of corners, nodes, leaves, etc. Typically, a function and threshold are applied to a node and the information is divided down the tree where, for instance, if the information is below a threshold, it goes left and right above a threshold until it ends up in a final cluster or class [33]. One DT technique is an ID3 algorithm that uses entropy to quantify data. The entropy is given below.

$$\textbf{Entropy}: \mathbf{H}(\mathbf{p_1}, \mathbf{p_2}, \dots \mathbf{p_s}) = \sum_{\mathbf{i=1}}^{\mathbf{s}} (\mathbf{p_i} \mathbf{log}\, (\mathbf{p_i})) \tag{8}$$

Where $(p_1, p_2, \dots p_s)$ represents the probabilities of the class labels.

Gini index is a metric of sample inequality. It has a value of 0 to 1. Gini value index 0 implies that the sample is completely homogeneous and all components are comparable, whereas Gini value index 1 implies maximum element inequality. It is the sum of each class's square probabilities. It is shown as,

$$\textbf{Gini index} = \mathbf{1} - \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{p_i^2} \tag{9}$$

4.1.5 Multilayer Perceptron (MLP)

A neural network is a sequence of algorithms that attempt to acknowledge fundamental interactions in a collection of information through a method that mimics the functioning of the human brain. Neural networks can adapt to altering inputs; therefore, the network produces the best possible result without redesigning the output requirements [32].

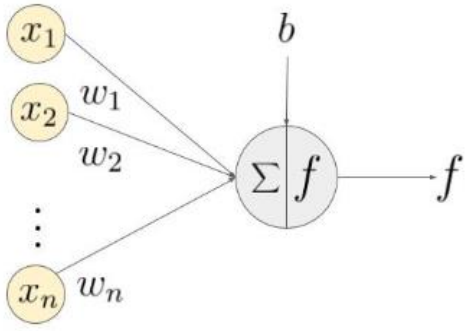

Figure 2: Neural Network Model [34]

$$\int (\boldsymbol{b} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i w_i}) \tag{10}$$

b = bias
x = neuron input
w = weights
n = number of incoming layer inputs
i = counter from 0 to n

4.1.6 Evaluation Metrics

Different metrics are used to evaluate the performance of our proposed model. These are mentioned below.

$$\textbf{Accuracy} = \frac{Tp+Tn}{Tp+Fp+Fn+Tn} \tag{11}$$

$$\textbf{Precision} = \frac{\text{Tp}}{\text{Tp+Fp}} \tag{12}$$

$$\boldsymbol{Recall} = \frac{Tp}{Tp+Fn} \tag{13}$$

$$\textbf{F1 Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{14}$$

5. Experiments and results

Once the data is clean, normalized and balanced, the data is loaded into Weka for analysis and comparison of various machine learning algorithms. Dataset is split as 30% for testing and 70% for training. Based on comprehensive experiments results are tabulated. Let's look into them one by one.

5.1 Experiment Results of Proposed Method

The experiment uses 50% percent of data containing "No Rain" class and 50% with "Rain" class data. Results are tabulated and shown in Table 3. Most of the algorithms have shown better results with 99% accuracy for both classed "No Rain" and "Rain" respectively, precision recall and F1 Measure. 50:50 Data Balancing Ratio For No Rain and Rain Class Results.

Table 3
Experiment Results of Proposed Method

No	ML Algorithm	Accuracy	No Rain Class – 50%					Rain Class – 50%				
			TP Rate	FP Rate	Precision	Recall	F1 Measure	TP Rate	FP Rate	Precision	Recall	F1 Measure
1	Naïve Bayes	95.76	0.03	0.96	0.96	0.95	0.95	0.96	0.04	0.95	0.96	0.95
2	SVM	99.15	0.99	0	0.99	0.99	0.99	0.99	0.01	0.99	0.99	0.99
3	KNN	98.76	0.97	0	0.99	0.97	0.98	0.99	0.02	0.97	0.99	0.98
4	DT	99.80	0.99	0	1	0.99	0.99	1	0	0.99	1	0.99
5	MLP	99.21	0.98	0	0.99	0.98	0.99	0.99	0	0.98	0.99	0.99

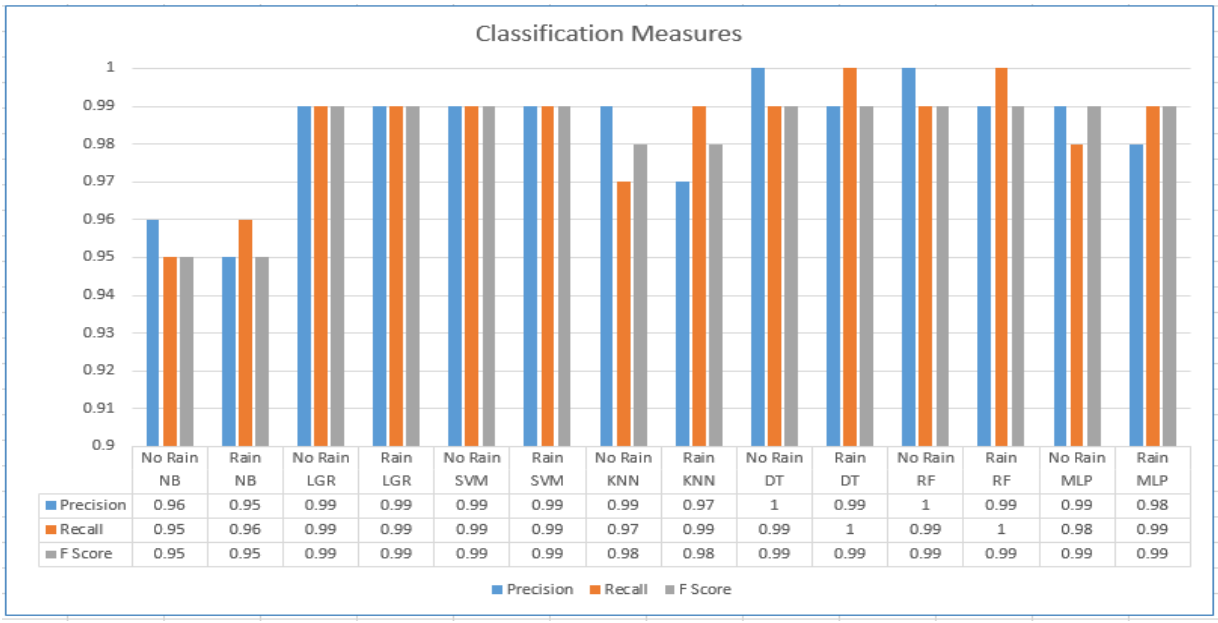

Figure 3: Proposed Model Experiment Result

5.2 Comparison of proposed technique with base paper (no rain class)

The Table 4 shows the comparison of our proposed method with the existing study based on "No Rain" class. The results based on precisions, Recall and F-1 measure shows that overall our proposed method improves on average 6%, 0.4% and 4% respectively.

Table 4
Comparison of Proposed technique with base paper (No Rain Class)

No	ML Algorithm	Proposed Method			Base Paper		
		Precision	Recall	F1 Measure	Precision	Recall	F1 Measure
1	Naïve Bayes	0.96	0.95	0.95	0.93	0.95	0.94
2	SVM	0.99	0.99	0.99	0.91	1	0.955
3	KNN	0.99	0.97	0.98	0.93	0.94	0.94
4	Decision Tree	1	0.99	0.99	0.93	0.98	0.95
5	MLP	0.99	0.98	0.99	0.93	0.99	0.96

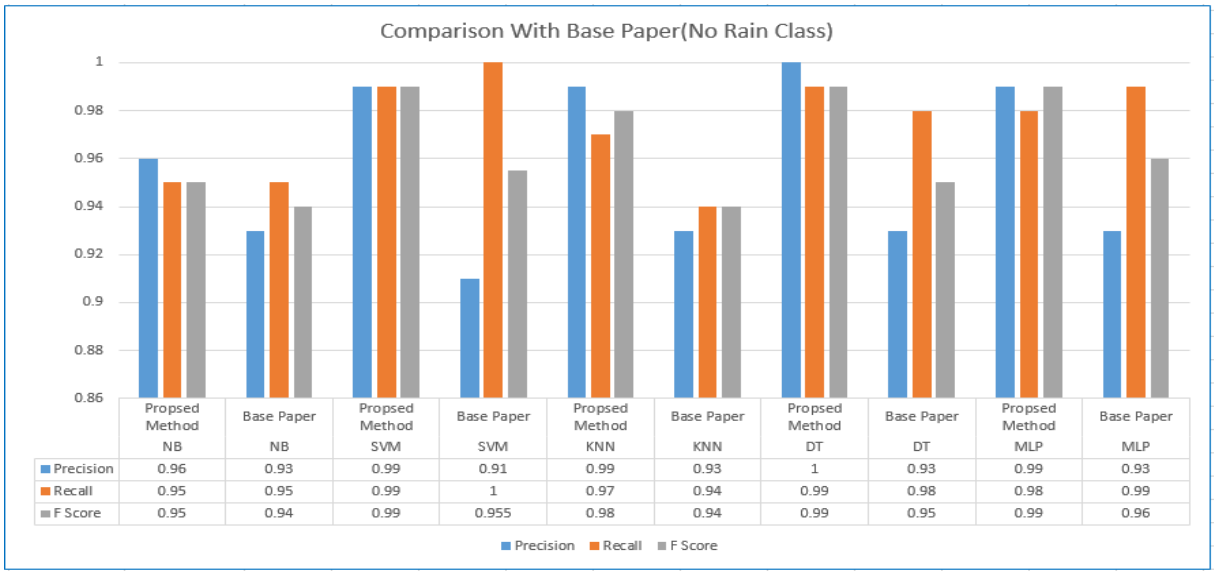

Figure 4: Proposed Method Comparison with Base Paper (No Rain Class)

5.3 Comparison of proposed technique with base paper (rain class)

Similarly in table 5 shows the comparison of our proposed method with the existing study based on "Rain" class. The results based on precisions, Recall and F-1 measure shows that overall our proposed method improves on average 58%, 76%, and 72% respectively.

Table 5
Comparison of Proposed technique with base paper (Rain Class)

No	ML Algorithm	Proposed Method			Base Paper		
		Precision	Recall	F1 Measure	Precision	Recall	F1 Measure
1	Naïve Bayes	0.96	0.95	0.95	0.36	0.28	0.31
2	SVM	0.99	0.99	0.99	0	0	0
3	KNN	0.99	0.97	0.98	0.36	0.31	0.33
4	Decision Tree	1	0.99	0.99	0.59	0.24	0.34
5	MLP	0.99	0.98	0.99	0.72	0.22	0.33

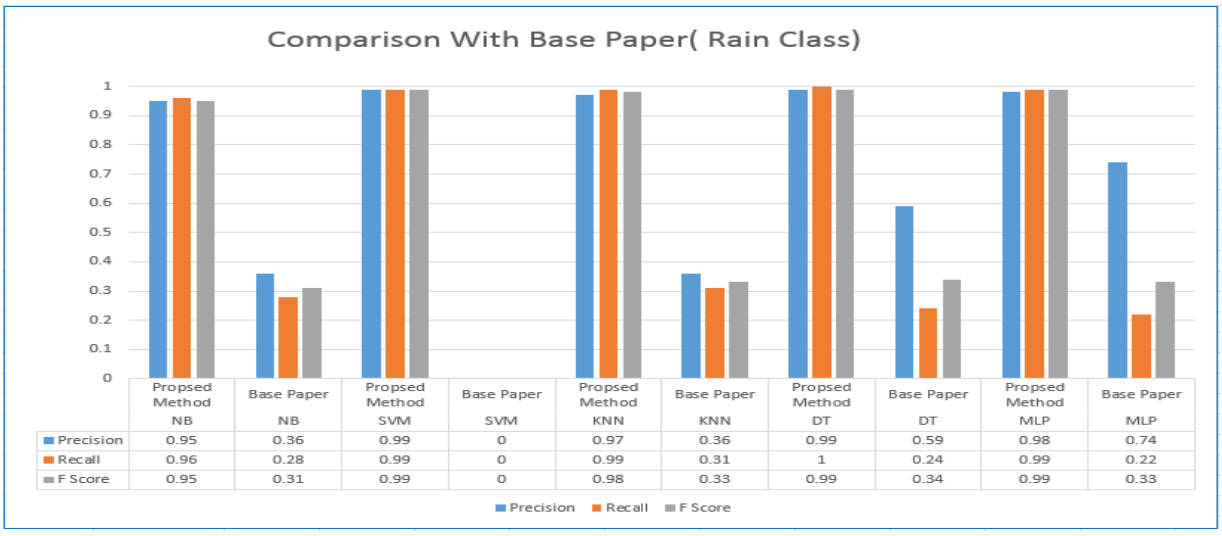

Figure 5: Proposed Method Comparison with Base Paper (Rain Class)

5.4 Critical Analysis

Shabib Aftab, Munir Ahmad [9] used data mining techniques for Lahore rain predication, experiments showed good results for a no-rain class in terms of precision, recall, and f-measures. However for rain class these techniques did not perform well and results were not accurate. So more robust method was needed to solve this problem. In order to improve results for rain class we proposed SMOTE which balance the data and then this balance data is pass to machine learning models to train and test the performance. The results based on precisions, Recall and F-1 measure shows that overall our proposed method improves on average 58%, 76%, and 72% respectively for rain class and precisions, Recall and F-1 measure was improved on average 6%, 0.4% and 4% respectively.

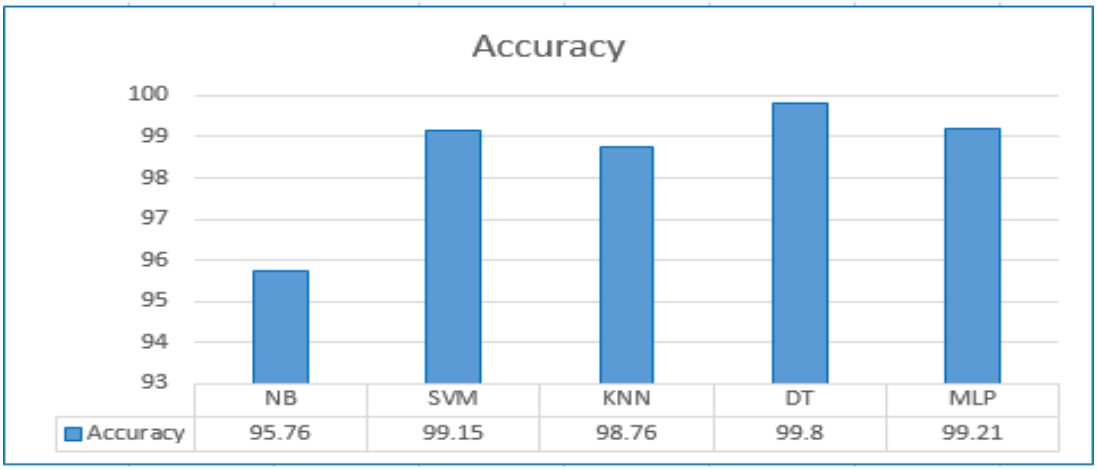

Figure 6: Comparison of Accuracies Algorithms Used in This Research

Similarly, figure 5 shows the comparison of accuracy achieved in this research using different machine learning algorithms. We achieved 95.76% accuracy using Naive Bayes (NB) algorithm. Similarly, the accuracy we achieved from KNN was 99.15%. 99.15%, 99.8%, and 99.21% accuracy were achieved using Support Vector Machine (SVM), Decision Tree (DT) and Multi-layer-perceptron (MLP).

6. Conclusion and future work

In this paper, we have performed rain predication for using seven Machine Learning Algorithms: Naive Bayes, Support Vector Machine (SVM), Multilayer Perceptron (MLP), K Nearest Neighbor and Decision Tree. For this purpose, we have used 12 years of time series data from December 2005 to November 2017.The dataset was also used by [9] for experiments and there study showed that the data they collected produced good results for one class which was no-rain class but rain class results were not good. The reason was the dataset was imbalance. So in this study data balancing method name SMOTE was used to balance the dataset and then performed preprocessing to clean the dataset and replace the missing value by mean. After replacing the missing values dataset is normalized using Z-score normalization to bring the data into one scale. After that dataset is divided into two sets one is training having 70% data and another set is testing having 30% data respectively.

Machine learning algorithms show that data balancing has improved the results. Our experiment results show that Decision Tree performed well for both the classes in terms of precision, recall and f1-scores. The proposed model can be recommended for the major cities in Pakistan for rain prediction in practice. In future we will use deep learning approach like LSTM to identify the behavior of weather time wise.

7. References

[1] Razaque, Abdul, Marzhan Abenova, Munif Alotaibi, Bandar Alotaibi, Hamoud Alshammari, Salim Hariri, and Aziz Alotaibi. (2022). Anomaly detection paradigm for multivariate time series data mining for healthcare. Applied Sciences 12, no. 17: 8902.

[2] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay (2017). A Comprehensive Survey of Data Mining Techniques on Time Series Data for Rainfall Prediction," vol. 11, no. 2, pp. 168–184.

[3] M. Ahmad, S. Aftab, and S. S. Muhammad. (2017). Machine Learning Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27.

[4] Razaque, A., Amsaad, F., Hariri, S., Almasri, M., Rizvi, S. S., & Frej, M. B. H. (2020). Enhanced grey risk assessment model for support of cloud service provider. IEEE Access, 8, 80812-80826.

[5] K. Lu and L. Wang. (2011). A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction. 2011 Fourth Int. Jt. Conf. Comput. Sci. Optim., pp. 1343-1346.

[6] W. C.L. and K.-W. Chau. (2012). Prediction of Rainfall Time Series Using Modular Soft Computing Methods. Eng. Appl. Artif. Intell., vol. 26, no. 852, pp. 1–37.

[7] V. B. Nikam and B. B. Meshram. (2013). Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach. 2013 Fifth Int. Conf. Comput. Intell. Model. Simul., pp. 132-136.

[8] Al-Qubaydhi, N., Alenezi, A., Alanazi, T., Senyor, A., Alanezi, N., Alotaibi, B., ... & Alotaibi, A. (2022). Detection of unauthorized unmanned aerial vehicles using YOLOv5 and transfer learning. Electronics, 11(17), 2669.

[9] Shabib Aftab, Munir Ahmad, Noureen Hameed, Muhammad Salman Bashir, Iftikhar Ali, Zahid Nawaz. (2018). Rainfall Prediction in Lahore City using Data Mining Techniques. Int. J. of Advanced Computer Science and Applications, Vol. 9, No. 4, pp. 254-260.

[10] M. Ahmad and S. Aftab. (2017). Analyzing the Performance of SVM for Polarity Detection with Different Datasets, Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36.

[11] M. Ahmad, S. Aftab, and S. S. Muhammad. (2017). Machine Learning Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27.

[12] M. Ahmad, S. Aftab, I. Ali, and N. Hameed. (2017). Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3.

[13] "http://ru8.rp5.ru/Weather_archive_in_Lahore." [Online]. Available: http://ru8.rp5.ru/Weather_archive_in_Lahore.

[14] C. Sivapragasam, S. Liong, and M. Pasha. (2001). Rainfall and runoff forecasting with SSA-SVM approach," J. Hydroinformatics, no. April 2016, pp. 141–152.

[15] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar. (2008). Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine," IEEE Trans. Knowl. Data Eng., vol. 20, no. 9, pp. 1264–1272.

[16] Krasnopolsky, Vladimir M., and Michael S. FoxRabinovitz. (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Networks. 19.2. pp. 122-134.

[17] Lai, Loi Lei, et al. (2004). Intelligent weather forecast."Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on. Vol. 7. IEEE.

[18] Radhika, Y., and M. Shashi. (2009). Atmospheric temperature prediction using support vector machines. "International Journal of Computer Theory and Engineering 1.1: 55.

[19] Antonio S., et al. (2002). Bayesian networks for probabilistic weather prediction."15th Eureopean Conference on Artificial Intelligence (ECAI).

[20] R. Usha Rani, T.K.Rama Krishna Rao, R. Kiran Kumar Reddy (2015). An Efficient Machine Learning Regression Model for Rainfall Prediction" International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 23, pp.25-30

[21] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," Expert Syst. Appl., vol. 85, pp. 169– 181.
[22] C. Zhang, W.-B. Chen, X. Chen, R. Tiwari, L. Yang and G. Warner, "A Multimodal Data Mining Framework for Revealing Common Sources of Spam Images", Journal of multimedia, vol. 4, no. 5. October, pp. 313-320.
[23] C. Li, M. Zhang, C. Xing and J. Hu. (2011). Survey and Review on Key Technologies of Column Oriented Database Systems. Computer Science, vol. 37, no. 12, pp. 1-8.
[24] M. Zhang. (2011). Application of Data Mining Technology in Digital Library", Journal of Computers, vol. 6, no. 4, pp. 761-768.
[25] C.-W. Shen, H.-C. Lee, C.-C. Chou and C.-C. Cheng. (2011). Data Mining the Data Processing Technologies for Inventory Management", Journal of Computers, vol. 6, no. 4, pp. 784-791.
[26] Z. Danping and D. Jin. (2011). The Data Mining of the Human Resources Data Warehouse in University Based on Association Rule", Journal of Computers, vol. 6, no. 1, pp. 139-146.
[27] J. Jiang, B. Guo, W. Mo and K. Fan. (2012). Block-Based Parallel Intra Prediction Scheme for HEVC", Journal of Multimedia, vol. 7, no. 4, pp. 289-294.
[28] S.-Y. Yang, C.-M. Chao, P.-Z. Chen and C.-Hao. (2011). SunIncremental Mining of Closed Sequential Patterns in Multiple Data Streams", Journal of Networks, vol. 6, no. 5, pp. 728-735.
[29] G.Vamsi Krishna (2015). An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis" International Journal of Computer Applications (0975 – 8887) Volume 120 – No.11, pp.26-29.
[30] Nele Verbiest, Enislay Ramento, Chris Cornelis, Francisco Herrera. (2014). Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection". Elsevier - Applied Soft Computing Journal. 511–517.
[31] Lina Guzman, DIRECTV, Data sampling improvement by developing SMOTE technique in SAS, Paper 3483-2015.
[32] Almiani, M., AbuGhazleh, A., Al-Rahayfeh, A., & Razaque, A. (2020). Cascaded hybrid intrusion detection model based on SOM and RBF neural networks. Concurrency and Computation: Practice and Experience, 32(21), e5233.