Prediction of weather using high-performance gradient boosting

V. Bibin Christopher*

Department of Computing Technologies,
SRM Institute of Science and Technology,
Kattankulathur Campus, Chengalpattu Dt.,
Tamilnadu, India
Email: bibinchrist85@gmail.com
*Corresponding author

R. Isaac Sajan

Department of Electronics and Communication Engineering,
Ponjesly College of Engineering,
Alamparai, Tamilnadu, India
Email: isaacsajanr.001@gmail.com

T.S. Akhila

Department of Electronics and Communication Engineering,
Mar Ephraem College of Engineering and Technology,
Lavuvilai, Marthandam, Tamilnadu, India
Email: akhilats31@gmail.com

M. Joselin Kavitha

Department of Electronics and Communication Engineering,
Marthandam College of Engineering and Technology,
Marthandam, Tamilnadu, India
Email: drjoselinkavitha@gmail.com

Abstract: Our weather prediction technology is imprecise despite its many new uses. Thus, demand exists to adopt a new method that eliminates the system's drawbacks and accurately projects rain. Existing machine learning methods use more RAM, are hard to trim, take a long time to compute, and are hard to use for time series predicting datasets. A high-performance gradient-boosting framework-based decision tree algorithm predicts rain. We used light gradient boosting machine (Light GBM), a leaf-wise method with best-fitting models that eliminates overfitting better than other decision tree algorithms. Predicting continuous goal variables is faster, more efficient, and uses less memory. Rain is Seattle's trademark. This study uses the Seattle dataset of daily weather from 1948 to 2017. The goal is to compute DATE, PRCP, TMAX, TMIN, and RAIN at each break and create a final forecast based on the sampled light BGM that is more accurate than other boosting algorithms.

Keywords: light gradient boosting machine; light GBM; leaf-wise algorithm; precipitation; PRCP; temperature maximum; TMAX; temperature minimum; TMIN; weather forecasting.

Reference to this paper should be made as follows: Christopher, V.B., Sajan, R.I., Akhila, T.S. and Kavitha, M.J. (2023) 'Prediction of weather using high-performance gradient boosting', *Int. J. Global Warming*, Vol. 31, No. 1, pp.30–41.

Biographical notes: V. Bibin Christopher received his BE in Electrical Engineering 2006. He received his ME and PhD in Computer Science and Engineering from Anna University, Chennai, India in 2009 and 2021. Currently, he is working as an Associate Professor in SRM Institute of Science and Technology, Chennai, India. His current research interests include wireless sensor networks, mobile computing, and network security.

R. Isaac Sajan received his BE, ME and PhD in Computer Science and Engineering from Anna University, Chennai, India in 2006, 2008 and 2021. Currently, he is working as the Vice-Principal and an Associate Professor in the Department of Electronics and Communication Engineering, Ponjesly College of Engineering, Nagercoil, India. His current research interests include wireless sensor networks, mobile computing, wireless communications in general and artificial intelligence.

T.S. Akhila received her BE in Electronics and Communication Engineering and ME in Embedded System Technologies from Anna University, India in 2009 and 2011, respectively. She is an Assistant Professor in Mar Ephraem College of Engineering and Technology, India. Her research interests are wireless sensor networks, wireless communication, VLSI design and embedded system.

M. Joselin Kavitha received her BE in Electronics and Communication Engineering and ME in Communication Systems from Anna University, India in 2008 and 2010, respectively. She is an Assistant Professor in Marthandam College of Engineering and Technology, India. Her research interests are wireless sensor networks, wireless communication and VLSI design.

1 Introduction

Rainfall is typically necessary for the productivity of agricultural outputs as well as industrial growth that is reliant on agriculture. Rainfall is one of the elements that has a role in determining the success of agricultural farming in India. The agricultural production of rain-fed crop regions in the nation has ramifications not only for the country's economy but also for its politics and society. In spite of the fact that competing with nature is a difficult endeavour, one may always make an effort to be cautious in order to avoid serious problems brought on by incorrect forecasts. Given that some weather occurrences have been known to cause loss of life, damage to infrastructure, and reductions in agricultural output, it is important to take weather seriously. Flooding and significant damage have been caused as a consequence of an unexpected or abrupt downpour that occurred within a short period of time. The flooding may cause harm to the crops. For this reason, it is very necessary to carry out accurate forecasting,

particularly in nations such as India that are reliant on the agricultural production. In addition, the accuracy with which weather forecasts are made is of the utmost importance for day-to-day well-being or living.

Agriculture has a significant role in the Indian economy, which is strongly reliant on the sector. In India, about 70% of the population relies on agricultural operations in some capacity, whether they do so directly or indirectly. Production in the agricultural sector has been the primary factor in determining India's gross domestic product (GDP), which in turn has been the primary factor in determining agriculture's direct or indirect effect on employment described by Krishna-Kumar et al (2004). In light of these requirements, the purpose of this study is to provide a solution to the challenging problem of hydro-meteorologists to accurately anticipate the occurrence of rainfall events using historical data that is readily accessible over a short period of time. The vast majority of meteorological prediction models include mathematical models that are quite difficult to compute and call for much prior experience or knowledge in this area (Charlton-Perez et al. 2015).

As a recent pattern has shown, the fast growth of computers and advancements in the field of information technology have led to a rise in the number of applications of computational intelligence in many parts of meteorology. The method known as computational intelligence or soft computing has shown a substantial change in the progression of new hybrid data-driven approaches in the modelling of a broad variety of meteorological data (Ikuta et al., 2021). Also, data-driven models are easy to grasp and need no previous knowledge or skill.

In addition, the performance might be improved by using many soft computing strategies, which include integrating human thinking and behaviour in a computer setting. Approaches to soft computing integrate efficient computational procedures that are encouraged by the inherent fuzziness, intuition, and familiarity of logical thinking and real-world ambiguity. KNN-based weather prediction has been more popular in recent years as the need for increasingly complex forecast models continues to rise. The use of KNN for weather prediction is an example of hybridisation, which refers to the coupling of many different soft computing techniques in order to accomplish a certain goal.

Approaches from rough set theory, neural networks, evolutionary algorithms, and fuzzy sets are often implemented in hybrid architectures because these methods are best suited to manage uncertainty in the real-world issues they are meant to solve (Madeira et al., 2021). There is a degree of haziness and unpredictability in practically every situation because the dynamics of the atmosphere need it.

The probability theory, the Dempster-Shafer theory, rough sets, fuzzy logic, and soft sets are some of the recent theories that are being utilised to deal with uncertain, imprecise, and fuzzy information respectively. With the use of KNN as a tool for knowledge discovery, this study investigates the application of an intelligent computing strategy to the problem of incorporating uncertainty and fuzziness into weather models. In this investigation, a method known as feature selection based on rough sets is combined with a neural network that simulates the reasoning and learning capabilities of humans. This combination is used to great advantage in order to improve prediction accuracy. The ambiguity and fuzziness of the rainfall dataset may be handled via rough set theory. In order to improve the accuracy of the rainfall forecast, intelligent methodologies are being integrated into the modelling process of a competitive and intelligent rainfall prediction model.

The demand for precise weather forecasting never decreases, regardless of how much time passes. Time series forecasting is the primary method used in weather forecasting. The prediction of future weather conditions using the historical values of weather patterns and models is referred to as time series forecasting. This kind of forecasting is used extensively in meteorology. Even if there is no assurance that this method would be successful, the scientist suggests using more recent technology in order to overcome the cyclical fluctuation analysis and the problems associated with seasonality (Kopitar et al., 2020). Because weather predicting has always remained a complex science for such a long period of time, using an algorithm that uses machine learning to handle time series forecasting is a technique that is acceptable. The precise forecasting of weather and rainfall will be more beneficial in increasing agricultural output, arranging holidays, preparing preliminary plans for water infrastructure, organising outdoor wedding receptions, constructing buildings, booking flights, and working on industrial projects. The process of predicting first relies on local measurements of the sky, the wind, and the temperature. These observations may be divided into two categories: surface-based observations and upper-air observations. Today, an assessment of the equations governing fluid mechanics and thermodynamics is included in the process of numerical weather prediction. This allows forecasters to determine how the weather and atmospheric conditions will evolve in the future. There are a few issues with these forecasts, such as the use of faulty data and the management of enormous datasets, both of which take extra processing effort. It is possible to find a solution to this problem via the use of improvised data-driven global weather forecast employing machine algorithms. The approaches that WE have developed perform better than those that are based on weather prediction models with a coarse resolution as well as current shortcomings that are caused by regression and classification issues.

2 Problem statement and problem solutions

Existing methods like ARIMA/State space needs data linearisation. It requires complex data pre-processing. There is a data loss during autocorrelation and partial autocorrelation. Data linearisation becomes more complex and less accurate with time-series data. FitNet depends mostly on past values, which might lead to overfitting, and in turn, it produces high MAPE values (Qasim, 2021). Other Decision tree problems are in need of more memory space, difficulty to prune, and high computational time.

Our proposed methodology is also one of the boosting algorithms which do not require more space or computational time. It is easy to prune, and it is the best solution to the overfitting problem faced by the existing methodologies. The light gradient boosting method can be used to predict rainfall. It is, based on decision algorithms, but it splits the tree according to the leaf in the tree with the best fit, whereas other decision tree algorithms split the tree based on depth or level basis rather than leaf basis. In which LGBM reduces the data loss, and subsequently gives good accuracy. Light GBM is faster in handling the large dataset than the XGBOOST and gradient BOOST algorithm. The computation time will be also reduced because the Light GBM uses tuning parameter for Best fit to beat the overfitting problems

3 Related work

A mathematical methodology Pawlak developed called the 'rough set approach' deals with ambiguity and incomplete information. RST does not need any prerequisite knowledge or further details about the data. According to Pawlak (1997), the mathematical foundation of rough set theory is the indiscernible relation, which states that every object in the universe set has some information, objects with the same values are imperceptible given the information available, and objects with different values are discernible. Widely used in the area of intelligent data analysis are rough set techniques. It is particularly well suited for parallel processing, identifying minimum datasets, providing workable approaches to uncover hidden information in data, and assessing the significance of the data. A decision rule collection of information is produced using rough computing, and the outcomes are easily read.

According to Pawlak and Skowron (2007a), rough set theory may be seen as a specific application of Frege's concept of vagueness, whereby imprecision is reflected in a set's border area.

According to Pawlak and Skowron (2007b), decision-making in dynamic and unpredictable environments is a typical difficulty in weather forecasting. This is supported by cognitive sciences and artificial intelligence. Rough sets have been utilised in a variety of applications, mostly for knowledge discovery in the areas of business, the arts, and meteorological concerns as described in Shen and Jensen (2007).

The support vector machine (SVM), artificial neural network (ANN), and a time series based recurrent neural network are the three different machine learning models that were used to predict the weather. It also discussed the procedures that were carried out in order to achieve the results. In order to make accurate forecasts of the weather, a RNN based on time series is combined with a linear SVC and a five-layered neural network. The outcomes of these models are analysed and compared on the basis of the Root Mean Squared Error, which is the difference between the values that were predicted and the values that were actually observed (Singh et al., 2019). Due to its impact on human existence worldwide, weather forecasting has drawn the interest of several researchers from diverse study groups. Many studies have been motivated to explore hidden hierarchical patterns in the large volume of weather dataset for weather forecasting as a result of the recent development of deep learning techniques, the widespread availability of massive weather observation data, and the advent of information and computer technology. This research examines deep learning methods for predicting the weather. The prediction performance of the recurrence neural network (RNN), conditional restricted Boltzmann machine (CRBM), and convolutional network (CN) models will be specifically compared in this research. The weather dataset provided by BMKG (Indonesian Agency for Meteorology, Climatology, and Geophysics), which was compiled from a number of weather stations in the Aceh area between 1973 and 2009, and the El-Nino Southern Oscillation (ENSO) dataset provided by International organisations like the National Weather Service Centre for Environmental Prediction Climate are used to test those models (NOAA). Each model's forecasting accuracy is assessed using the Frobenius norm (Salman et al., 2015).

4 System model

We A significant test in the time series forecasting field is to acquire sensibly precise forecasts of future information from breaking down records. A productive option in contrast to utilising multiple forecasting techniques is to enable the single forecasting ensemble learning method using a decent prediction algorithm (Mehta et al., 1996). Various examinations suggest a single forecasting methodology because it concentrates mainly on eliminating the most frequent problem, which lessens the forecasting blunders in the model. In this paper, we propose an ensemble technique which, specifically strengthens a portion of the basic forecasting models rather than combining different machine learning algorithms. On each time series, the segment models are progressively positioned according to their past forecasting exactness, and we strengthen the forecasting method by high positioned models like light GBM. In general ensemble method, is a machine learning technique that blends several base models to one best predictive model. There is numerous boosting algorithm in ensemble learning, but each one is based on decision tree model.

Light gradient boosted machine, or LightGBM gives a proficient and compelling execution of the gradient boosting calculation. LightGBM broadens the gradient boosting calculation, which is done by adding a sort of program which includes choice just as zeroing in on boosting models with larger gradients. This can bring an exceptional speedup of preparing and working on predictive execution. In that role, LightGBM has become an accepted calculation for machine learning rivalries when working with even information for regression and classification predictive modelling tasks. Hence, it possesses a portion of the fault for the expanded prominence and more extensive selection of gradient boosting strategies as a rule, alongside extreme gradient boosting (XGBoost) (Wang et al., 2017).

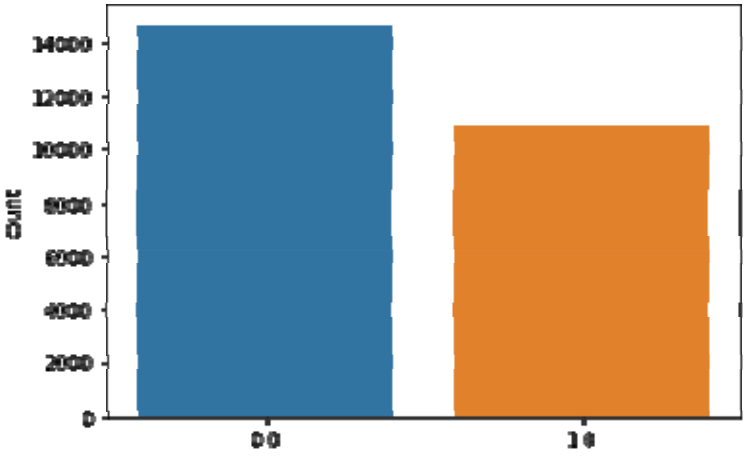
Light GBM is an emerging open-source technique in ensemble learning. It is considered one of the best prediction algorithms. So far, no one has used this algorithm to predict weather forecasting. Light GBM framework is to predict the rainfall prediction, which can be explored as one of the best-boosting methods based on decision tree algorithms. This is used for the process of ranking, classification, prediction, visualisation, and many other artificial intelligence tasks (Ke et al., 2017).

Light GBM is a gradient-boosting decision tree algorithm introduced by Microsoft in the year 2016 (Ke et al., 2017). Many scientists have been researching this algorithm to develop and explore the Light GBM for many applications. It is a histogram-based model used to fasten the training and testing process, reduce memory usage, simplify data processing, reduce computational time, optimise parallel learning, easy prune and provide accurate solutions to overfitting problems usually faced by decision tree algorithms. The light Gradient boosting machine (light GBM) uses both weak and strong learning processes depending upon the volume of the dataset called small and big gradients (gi). The leaf-wise strategy is used while splitting the dataset in my algorithm, and while the other boosting algorithm prefers level-wise tree growth that leads to overfitting problems. Overfitting happens during the learning phase, where it develops an inference, which reduces training set error at the cost of an increased test set error, which is avoided by the pre-pruning and post-pruning processes. The existing models are difficult to prune, but my proposed technique uses tuning parameters like num_leaves, min_data_in_leaf, max_depth categorised as the best fit. To enable speed a faster bagging_fraction, feature_fraction, max_bin parameters are used that are computationally inexpensive.

The variance gain Vj(d) is calculated over the subset |AUB| with the remaining set AC. The larger gradient is assigned as A and the smaller gradient as B. The d in the variance gain is the data splitting. The splitting of data is estimated for optimal gain in variance with the coefficient is used to normalise the gradients

$$V_j^*(d) = \frac{1}{n}\left(\frac{\left(\sum x_i A_l g_i + \frac{1-\alpha}{b}\sum x_i \in B_l g_i\right)^2}{n_r^j(d)}\right) + \left(\frac{\left(\sum x_i A_r g_i + \frac{1-\alpha}{b}\sum x_i \in B_r g_i\right)^2}{n_r^j(d)}\right)$$

Figure 1 Data visualisation to understand the dataset (see online version for colours)



- Stage 1: Dataset collection

  Alibaba reference datasets on Tianchwe of Seattle weather dataset contains records of rainfall patterns from 1-1-1948 to 31-12-2017.

- Stage 2: Data pre-preparation

  It utilises the technique known as date pre-processing treatment that handles the missing observations in the dataset either by dropping it or by correlation over features.

- Stage 3: Data visualisation

  Data visualisation is used to visualise the data to understand the dataset involved in classification of rain dataset as shown in Figure 1 and Figure 2 shows the correlation between TMAX vs. Rain. Figure 3 shows box-plot distribution diagram to determine whether TMAX vs. rain is normally distributed or not.

Figure 2 TMAX vs. rain (differs when rain and when it does not) (see online version for colours)

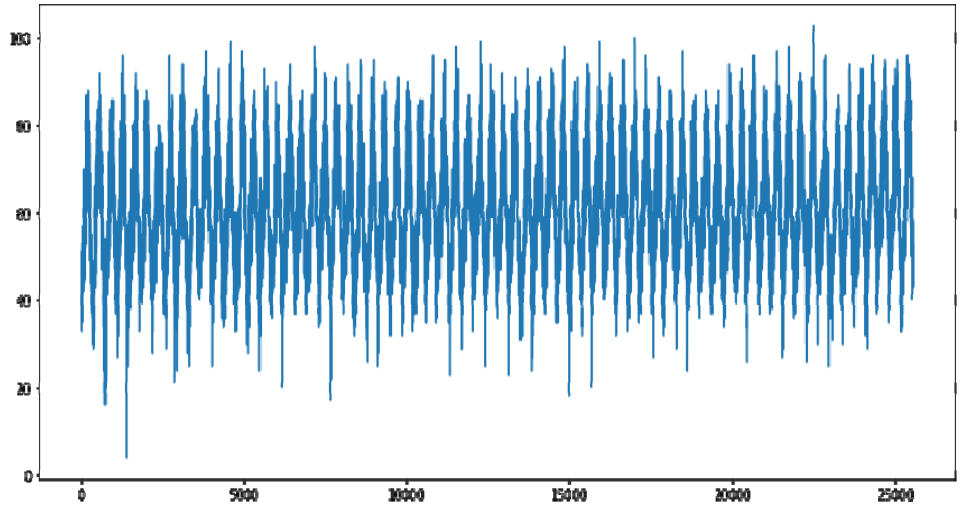
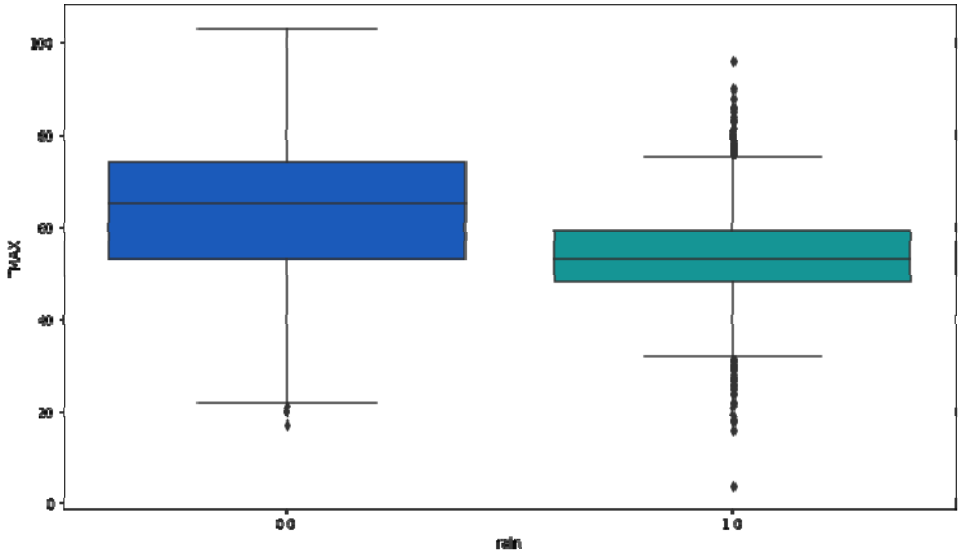


Figure 3 TMAX vs. rain (see online version for colours)

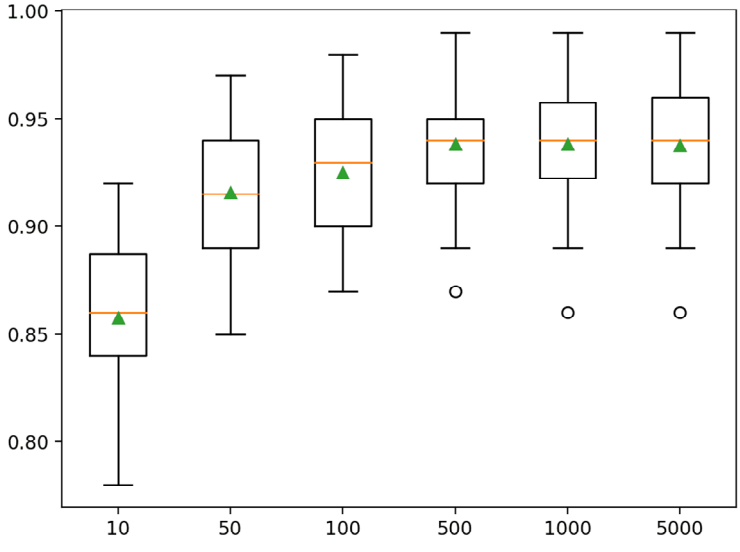


5 Stage: model training

Apply the regression through the GBDT technique by training the dataset inhibiting the stratified random method. Stratified random sampling is a strategy for sampling that includes the division of a dataset into more modest sub-bunches known as strata. In stratified random sampling or definition, the strata are framed dependent on an

individual's common attributes or qualities like pay or instructive achievement. A higher number of data's in the dataset to train and allow the model to perform better.

Figure 4 LightGBM ensemble size vs. classification accuracy (see online version for colours)



6 Performance evaluation

The performance of prediction of weather using high-performance gradient boosting has been evaluated, and the graphical representation of its accuracy is based on the learning rate, ensemble size, tree depth vs. classification is shown in Figures 4, 5 and 6.

Figure 5 Lightgbm ensemble tree depth vs. classification accuracy

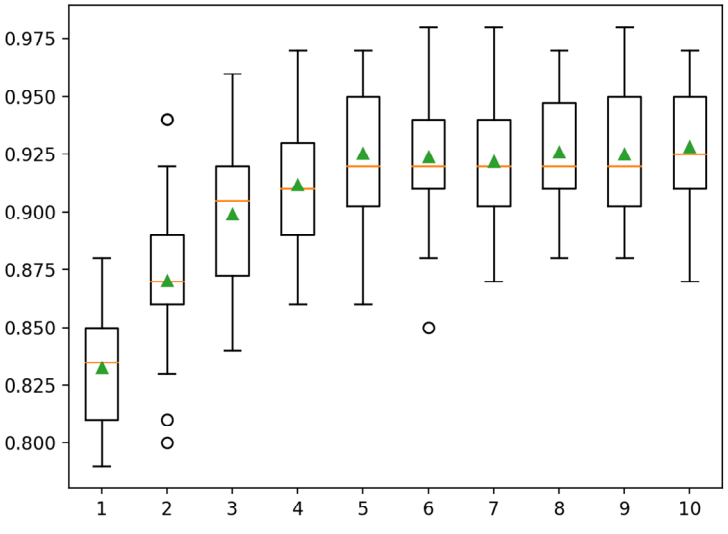
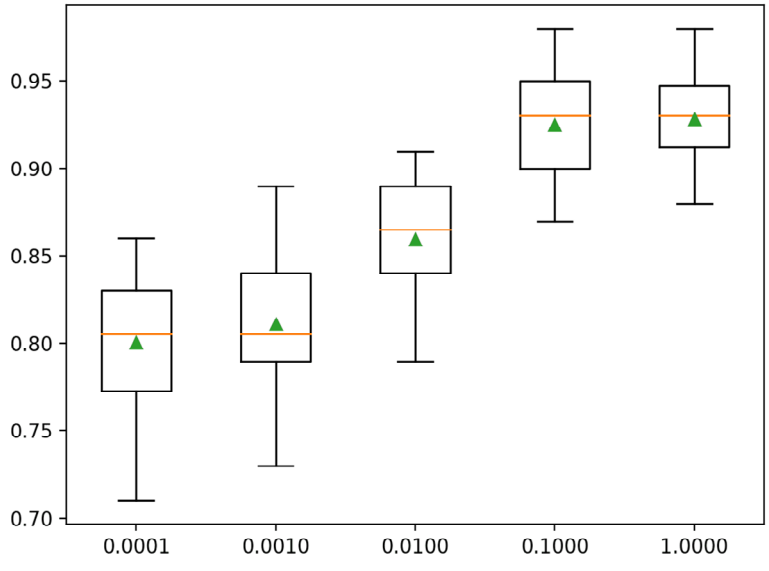


Table 1 Accuracy score vs. AUC score vs. execution time

	Accuracy score	*AUC score*	*Execution time*
Light GBM	0.961501	0.964492	00:00:00.283759
Xgboost	0.861398	0.764884	00:00:02.047220

Figure 6 Lightgbm learning rate vs. classification accuracy (see online version for colours)



7 Commercial and social value of our project

- Weather forecasting is important in many day-to-day activities like planning daily routines, official tours, or planning vacations.
- It is very useful for military officials in planning their activities.
- It is very useful for the aircraft official for landing and take-off.
- Even it is mainly useful for agriculture based on that they can cultivate the cops.

8 Conclusions

A High exhibition gradient boosting system-based decision tree calculation is utilised for foreseeing the precipitation. Light GBM is a leaf-wise calculation with best-fitting models, which breaks up the overfitting issue than the other decision tree algorithms. Also, it truly has quicker preparing speed and higher productivity, and its foreseeing consistent objective factors requires lower memory utilisation. Seattle is generally well known for is the way frequently it downpours. This paper utilises the Seattle dataset that contains total records of everyday precipitation designs from January 1, 1948, to

December 12, 2017. Our point is to ascertain DATE, PRCP, TMAX, TMIN, RAIN at each split, and make the last indicator dependent on the accumulated consequences of the examined Light BGM that gives better exactness contrasted with other existing boosting algorithms. By comparing LightGBM with XGBoost gathering learning procedures by applying both the algorithms to a dataset and afterward looking at the presentation. Here we are utilising a dataset that contains the data about weather from high rainfall places. The dataset comprises more than 32561 perceptions and 14 features. There has been just a slight expansion inexactness and AUC score by applying Light GBM over XGBOOST yet there is a critical contrast in the execution time for the preparation methodology. Light GBM is very nearly multiple times quicker than XGBOOST and is an improved methodology when managing enormous datasets. This ends up being a great benefit when you are dealing with huge datasets in restricted time rivalries.

Acknowledgements

We would like to show our gratitude to our institution for sharing their pearls of wisdom with us during the course of this research work. We are also immensely grateful to the well-wishers for their comments on an early version of the manuscript, although any errors are own and should not tarnish the reputations of these esteemed individuals.

References

Charlton-Perez, C., Cloke, H.L. and Ghelli, A. (2015) 'Rainfall: high-resolution observation and prediction', *Meteorological Applications*, Vol. 22, No. 1, pp.1–2.

Ikuta, Y., Fujita, T., Ota, Y. and Honda, Y. (2021) 'Variational data assimilation system for operational regional models at Japan meteorological agency', *Journal of the Meteorological Society of Japan*, Ser. II, Vol. 99, No. 6, pp.1563–1592.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W. et al. (2017) 'Lightgbm: a highly efficient gradient boosting decision tree', *Advances in Neural Information Processing Systems*, Vol. 30, No. 1, pp.1–8.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. and Stiglic, G. (2020) 'Early detection of type 2 diabetes mellitus using machine learning-based prediction models', *Scientific Reports*, Vol. 10, No. 1, pp.1–12.

Krishna Kumar, K., Rupa Kumar, K., Ashrit, R. G., Deshpande, N.R. and Hansen, J.W. (2004) 'Climate impacts on Indian agriculture', *International Journal of Climatology*, Vol. 24, No. 11, pp.1375–1393.

Madeira, B.C., Tasci, T. and Celebi, N. (2021) 'Prediction of student performance using rough set theory and backpropagation neural networks', *European Scientific Journal ESJ*, Vol. 17, No. 7, pp.1–15.

Mehta, M., Agrawal, R. and Rissanen, J. (1996) 'SLIQ: a fast scalable classifier for data mining', in *International Conference On Extending Database Technology,* pp.18–32, Springer, Berlin, Heidelberg.

Pawlak, Z. (1997) 'Rough set approach to knowledge-based decision support', *European Journal of Operational Research*, Vol. 99, No. 1, pp.48–57.

Pawlak, Z. and Skowron, A. (2007) 'Rough sets: some extensions', *Information Sciences*, Vol. 177, No. 1, pp.28–40.

Pawlak, Z. and Skowron, A. (2007) 'Rudiments of rough sets', *Information Sciences*, Vol. 177, No. 1, pp.3–27.

Qasim, T. (2021) 'Estimating and forecasting meat prices in Pakistan: a comparative study of ARIMA, GARCH and state space ARIMA models', *Pakistan Social Sciences Review*, Vol. 5, No. 3, pp.151–176.

Salman, A.G., Kanigoro, B. and Heryadi, Y. (2015) 'Weather forecasting using deep learning techniques', in *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp.281–285.

Shen, Q. and Jensen, R. (2007) 'Rough sets, their extensions and applications', *International Journal of Automation and Computing*, Vol. 4, No. 3, pp.217–228.

Singh, S., Kaushik, M., Gupta, A. and Malviya, A.K. (2019) 'Weather forecasting using machine learning techniques', in Proceedings of *2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*.

Wang, D., Zhang, Y. and Zhao, Y. (2017) 'LightGBM: an effective miRNA classification method in breast cancer patients', in Proceedings of the *2017 International Conference on Computational Biology and Bioinformatics*, pp.7–11.