Cloudburst Prediction In Northern India Region Using Machine Learning And Deep Learning

A THESIS SUBMITTED TO
DEFENCE INSTITUTE OF ADVANCED TECHNOLOGY , PUNE
FOR THE SEMESTER THREE EVALUATION (2021-2023) OF
MASTER OF TECHNOLOGY
IN
MODELLING AND SIMULATION

BY
KM SONIKA
(Registration No. 21-14-10)

UNDER THE SUPERVISION OF
Dr. Bharath Ramkrishna & Dr. Manmeet Singh

SCHOOL OF COMPUTER ENGINEERING
&
MATHEMATICAL SCIENCES
DEFENCE INSTITUTE OF TECHNOLOGY(DIAT), PUNE, INDIA

DEC 2022

Dedicated

To each and everyone who supported and motivated me.

CERTIFICATE

This is to certify that the Thesis entitled "**Cloudburst Prediction In Northern India Region Using Machine Learning And Deep Learning** " submitted to Defence Institute of Advanced Technology, Girinagar, for the award of the degree of **Master of Technology**, is the bonafide research work done by **Ms. Km Sonika** under my supervision. The contents of this thesis have not been submitted elsewhere for the award of any degree.

Dr. Manmeet Singh
External Supervisor
Scientist 'C'
IITM, Pune

Dr. Bharath Ramkrishna
Internal Supervisor
Professor
School of Computer Engineering
& Mathematical Sciences
Defence Institute of Advanced
Technology
Girinagar, Pune, INDIA

DECLARATION

This is to certify that the work presented in the Thesis entitled **"Cloudburst Prediction In Northern India Region Using Machine Learning And Deep Learning"**, is a bonafide work done by me under the supervision of **Dr. Manmeet Singh and Dr. Bharath Ramkrishna** and has not been submitted elsewhere for the award of any degree.

Date:____________

Place:____________

Name: Km Sonika
Roll No.: 21-14-10
Department. School of Computer Engineering & Mathematical Science
Defence Institute of Advanced Technology
Girinagar, Pune

COUNTERSIGNED

Dr. Manmeet Singh
Scientist C
Indian Institute of Tropical Meteorology(IITM), Pune

Dr. Bharath Ramkrishna
Professor
Defence Institute of Advanced Technology(DIAT)
Girinagar, Pune

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to **Dr. Bharath Ramkrishna**, School of Computer Engineering & Mathematical Sciences, for motivating me all through this work. I would like to thank for his timely guidance and support.

I would like to express my sincere gratitude to **Dr. Manmeet Singh**, Scientist 'C', Indian Institute of Tropical Meteorology, Pune for providing me the opportunity to take up this project and more importantly, his precious guidance and motivation, which helped me to conquern the difficulties in research and development of this project.

Above all, I offer my sincere thanks and gratitude to the merciful God Almighty for the providence and grace, He has given through this attempt.

ABSTRACT

Many weather and rain specialists have found it to be a life-threatening issue to predict severe rain storms. Events involving torrential rain are unpredictable by their very nature. Because they last just a short time and only affect a small region of land, strong rains are particularly difficult for meteorological specialists to anticipate. It is difficult to predict where and when the torrential rains that fall in and around the southern point of the Himalayas in India will occur. Since the Himalayas are home to the majority of the recorded severe rains, observations are few. As soon as the effects on life and property loss in ecosystems downstream are known, the majority of these events are recorded. Additionally, they frequently involve flash floods.Nothing further is known about these events, other than the India Meteorology Department (IMD) criterion of 100 mm/h or above precipitation over a geographic range of roughly 20–30 km2. The valley folds in the southern tip of the Indian Himalayas, between 1000 m and 2500 m elevation, are where the majority of torrential rain events take place. Convection has been proven to be the primary cause of heavy rain events, followed by a topographically locked system, with the exception of some of the large-scale flows demonstrated in some studies. This study's main goal was to use satellite observations to track the Uttarakhand flood disaster. In this work, we offer a cloud burst prediction model that uses deep learning and machine learning approaches to forecast the occurrence of cloud bursts at a location. For the heavy downpours that happened in Uttarakhand, India, over the previous two decades, we gathered data and created a model. This forecasting model will be a useful tool for streamlining decision-making and setting up early warning systems in the event of a cloudburst.

Contents

List of Figures

NOMENCLATURE

JAXA	Japan Aerospace Exploration Agency
ERA5	Fifth Generation ECMWF Atmospheric Reanalysis
AUC	Area Under Curve
U Wind	Horizontal Component Of Wind
V Wind	Vertical Component Of Wind
ECMWF	European Centre for Medium-Range Weather Forecasts

Chapter 1

Introduction and Literature Review

1.1 Introduction

1.1.1 Definition

A torrential downpour, a sudden, very heavy rain, usually localized and of short duration. Most so-called torrential rains are associated with thunderstorms [9]. These storms have violent gusts of wind that can prevent condensed raindrops from falling to the ground. A large amount of water can therefore accumulate at a high level, and all this water will fall at once when the upward flow slows down [9].Torrential rains are especially common in mountainous areas. This is probably because warm air currents from thunderstorms tend to flow along mountain slopes [3]. The impact of heavy rainfall is particularly pronounced on mountain slopes, with flooding concentrated in valleys and gorges. Heavy rains in mountains cause sudden and devastating floods. The intensity of precipitation during the most intense torrential rains can only be guessed.With a steep and erratic slope of degrees, the Himalayan terrain provides the perfect platform for heavy rain events that can cause degree flash floods and landslides [3]. Predicting the location, scale and magnitude of such catastrophic events remains a challenge.

Heavy rains occur during the monsoon season, but in the Himalayan region, pre-monsoon showers can also occur. Each year, this causes enormous loss of life, property, infrastructure, agricultural land and other facilities [7].Natural climate change plays a much more important role in causing these extreme events in the Himalayas [8]. The major roles of climate change, telecommunications, primary atmospheric circulation, and land use/land cover change are the main driving forces behind these extreme events [8]. As per [4] the low-level convergence of southeasterlies and northeasterlies along the foothills coupled with vertical shear in wind and orographic uplifting leading to a short-lived, intensely precipitating convective storm (cloudburst).

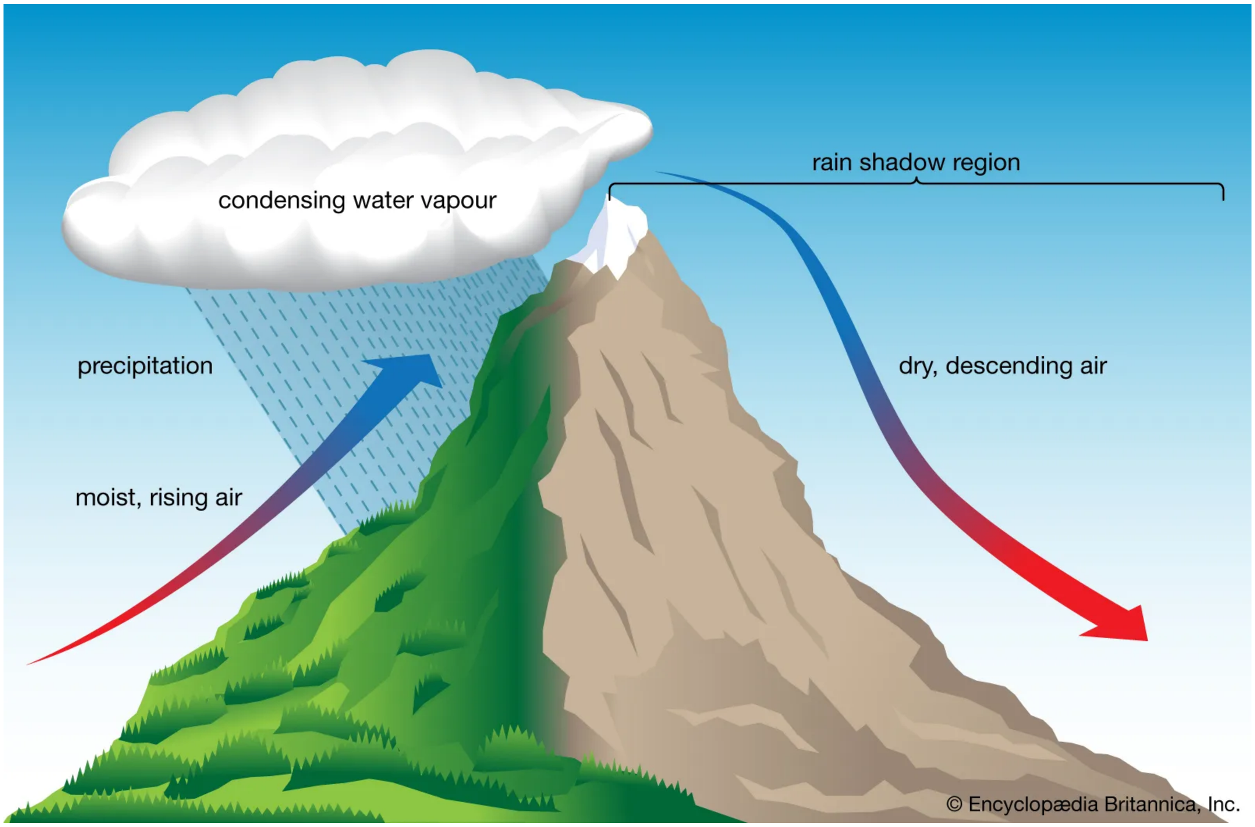
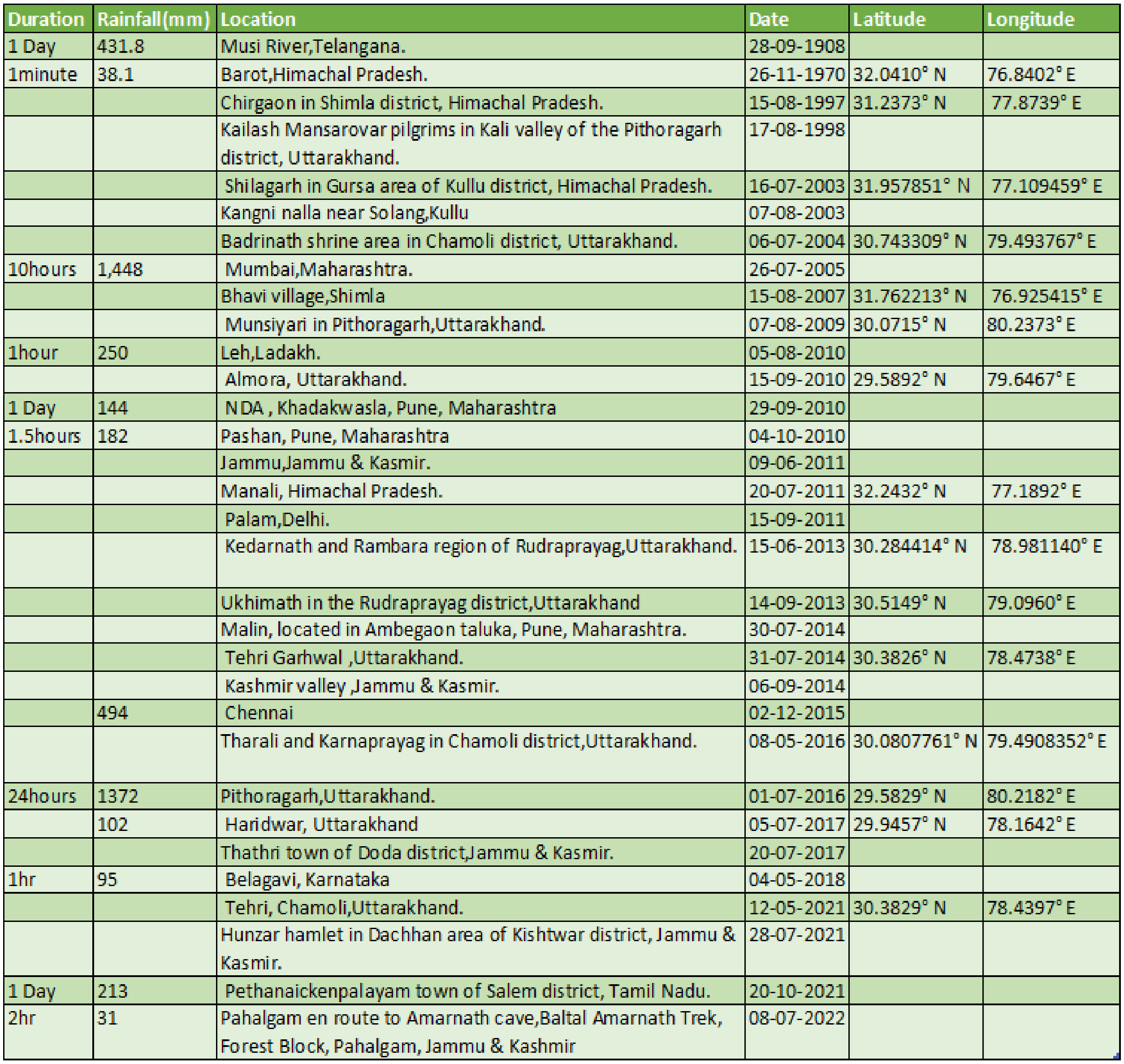


Figure 1.1: Orographically locked systems of Himalayan mountains leading to cloudburst

Figure 1.1 shows that cloudburst events are convectively triggered followed by orographically locked systems Himalayan mountains [3]. Previous disasters show that increasing showers and associated flash floods, debris flows and landslides are important sources of damage and destruction [7]. It is not yet confirmed when and in which region the same event will occur. Climate change due to global warming is the main concern for these extreme events [7].

According to Indian Meteorological Department (IMD), torrential rainfall occurs when high intensity (¿100 mm/h) precipitation falls over a small area in a short period of time and the same area does not exceed 20-30 km^2 [7]. It represents cumulonimbus convection under conditions of deep and rapid dynamic uplift with pronounced wet thermodynamic instability and steep topography [7]. It occurs mainly during the monsoon season from late June to early September and is caused by strong convective currents and terrain with steep slopes leading to flash floods and landslides in the Himalayan region [7].It attacks randomly and at lightning speed, and generally lasts for a limited amount of time, after which it leaves a trail of devastation.



Duration	Rainfall(mm)	Location	Date	Latitude	Longitude
1 Day	431.8	Musi River,Telangana.	28-09-1908		
1minute	38.1	Barot,Himachal Pradesh.	26-11-1970	32.0410° N	76.8402° E
		Chirgaon in Shimla district, Himachal Pradesh.	15-08-1997	31.2373° N	77.8739° E
		Kailash Mansarovar pilgrims in Kali valley of the Pithoragarh district, Uttarakhand.	17-08-1998		
		Shilagarh in Gursa area of Kullu district, Himachal Pradesh.	16-07-2003	31.957851° N	77.109459° E
		Kangni nalla near Solang,Kullu	07-08-2003		
		Badrinath shrine area in Chamoli district, Uttarakhand.	06-07-2004	30.743309° N	79.493767° E
10hours	1,448	Mumbai,Maharashtra.	26-07-2005		
		Bhavi village,Shimla	15-08-2007	31.762213° N	76.925415° E
		Munsiyari in Pithoragarh,Uttarakhand.	07-08-2009	30.0715° N	80.2373° E
1hour	250	Leh,Ladakh.	05-08-2010		
		Almora, Uttarakhand.	15-09-2010	29.5892° N	79.6467° E
1 Day	144	NDA , Khadakwasla, Pune, Maharashtra	29-09-2010		
1.5hours	182	Pashan, Pune, Maharashtra	04-10-2010		
		Jammu,Jammu & Kasmir.	09-06-2011		
		Manali, Himachal Pradesh.	20-07-2011	32.2432° N	77.1892° E
		Palam,Delhi.	15-09-2011		
		Kedarnath and Rambara region of Rudraprayag,Uttarakhand.	15-06-2013	30.284414° N	78.981140° E
		Ukhimath in the Rudraprayag district,Uttarakhand	14-09-2013	30.5149° N	79.0960° E
		Malin, located in Ambegaon taluka, Pune, Maharashtra.	30-07-2014		
		Tehri Garhwal ,Uttarakhand.	31-07-2014	30.3826° N	78.4738° E
		Kashmir valley ,Jammu & Kasmir.	06-09-2014		
	494	Chennai	02-12-2015		
		Tharali and Karnaprayag in Chamoli district,Uttarakhand.	08-05-2016	30.0807761° N	79.4908352° E
24hours	1372	Pithoragarh,Uttarakhand.	01-07-2016	29.5829° N	80.2182° E
	102	Haridwar, Uttarakhand	05-07-2017	29.9457° N	78.1642° E
		Thathri town of Doda district,Jammu & Kasmir.	20-07-2017		
1hr	95	Belagavi, Karnataka	04-05-2018		
		Tehri, Chamoli,Uttarakhand.	12-05-2021	30.3829° N	78.4397° E
		Hunzar hamlet in Dachhan area of Kishtwar district, Jammu & Kasmir.	28-07-2021		
1 Day	213	Pethanaickenpalayam town of Salem district, Tamil Nadu.	20-10-2021		
2hr	31	Pahalgam en route to Amarnath cave,Baltal Amarnath Trek, Forest Block, Pahalgam, Jammu & Kashmir	08-07-2022		

Figure 1.2: List past occurring cloudburst events in India with their locations [4]

Figure 1.2 shows the list past occurring cloudburst events with their locations and i clearly shows it occurs more frequently mainly with orographically locked systems like Himalayan and Western Ghats.

1.2 Literature Review

Kishan Rawat, Smruti Sahu, Sudhir Singh, and Anil Kumar Mishra [8] utilize satellite observations to follow the flood situation in Uttarakhand. This very severe rainfall in the state primarily affected the Nainital and Almora areas. At 11:00 am on October 18, Nainital recorded the day's highest rainfall total of about 21.51 mm (UTC) [8]. On October 18, 2021, the Nainital district received more than 300 mm of rain overall. According to the observation, the rapid cloud burst near Ramgarh in the Nainital district is what caused this significant rainfall. A flash flood also resulted from this abrupt, intense rainfall in the study region. After causing a flash flood in the Uttarakhand province's Nainital area, Assam has seen an increase in precipitation [8].

A.P. Dimri and A. Chevuturi and D. Niyogi and R.J. Thayyen and K. Ray and S.N. Tripathi and A.K. Pandey and U.C. Mohanty [4] synthesizes the available information and research on cloudburst events and tries to define it based on associated dynamics, thermodynamics and physical processes leading to a cloudburst event. They did analysis on characterizations and impacts of cloudburst leading from precipitation, dynamical, thermodynamical, large scale forcing to orographical forcing,followed geomorphology impacts are intertwined to present comprehensive portray of it [4]. Most of the cloudburst events are seen occurring in the elevation range of 1000 m to 2500 m within the valley folds of the southern rim of the Indian Himalayas. Apart from some of the large scale flow shown by few of the studies, it is found that cloudburst events are convectively triggered followed by orographically locked systems [4]. These intertwined mechanisms lead cloudburst events to form. Amiss of any one of these mechanisms will not lead the cloudburst mechanism to form. These interactions in the this paper established the vagaries associated with the cloudburst events [4].

Pranab Das [2] covers the floods and landslides that happened in the state between June 14 and June 17 of 2013 as a result of cloud bursts and significant rainfall. Landslides are highly common in Uttarakhand, and about three-fourths of the state's entire land area is in a region with a severe to high risk of landslides [2]. The disaster was caused by a cloud burst near Rambara in the Rudraprayag region, and nearly all of the major rivers, but especially the Mandakini and the Alaknanda, swelled as a result of torrential and prolonged rain. The region's overflowing water reservoirs were the catalyst for the event's intensity. Due to flooding and a landslide, almost 10 percent of Mandakini River's upper catchment area was destroyed. In the state's 80-year history, this was the deadliest threat [2]. The study reflects on nature's superiority over human potentials and promotes a harmonious relationship between man and nature, particularly in environments that are delicate and delicately balanced. It recommends using the Himalayan resources wisely and employing environmentally friendly methods for putting the region's development strategy into action [2]. Their investigation demonstrates that the main risk in Uttarakhand is cloudburst. Himalaya are a natural phenomenon that cannot be avoided, but if suitable precautions are adopted, disasters can be reduced. Additionally, extensive tree planting on the degraded terrain will lessen the severity of dangers [2].

Sivagami, M. and Radha, P. and Balasundaram, Ananthakrishnan [10] describe a cloudburst prediction model that makes use of deep learning methods to forecast when one will occur in a certain place. The model was created by the authors after they gathered data on the cloudburst incidents that took place in the Indian State of Uttarakhand over the previous ten years. Long Short Term Memory (LSTM) and Gated Recurrent Unit time series sequence models were used in the experiments (GRU) [10]. The crucial elements that are provided as input to these sequence models have been extracted using Predictive Power Score (PPS). When comparing GRU-based models to other sequence models, the findings are encouraging. The performance of sequence models has been evaluated in terms of loss function and accuracy [10].

Pranab Kr. Das [3] did examination of the Cloud Burst episodes in the Garwal-Kumaon Himalaya from a geographic perspective, with special attention to the influence of high altitude lakes and glacier melt water. Massive downpours known as ”Cloud Burst” that accompany landslides, flash floods, and earth flows can be devastating in mountainous places. The incidence of this occurrence has, however, greatly grown in recent years. The quicker evaporation rate from high altitude glacial lakes is one of the major factors contributing to frequent cloud bursts [3].Glacier melts water and lake comes direct contact with clouds due to higher altitude, create favorable condition of cloud burst in high altitude areas of Uttarakhand [3].

Chapter 2

Data Analysis and Data Preprocessing

2.1 Data Source

2.1.1 ERA5

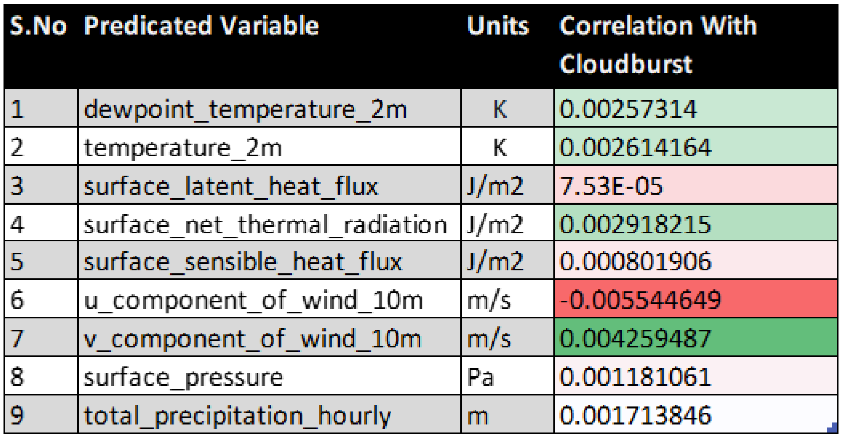
ERA5 is a global atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasting (ECMWF). It provides hourly data on many different atmospheric variables such as temperature, pressure, wind speed and direction, and precipitation. ERA5 is the fifth generation of the ECMWF's reanalysis datasets, and it is the most advanced and accurate reanalysis dataset available. It covers the period from 1979 to the present day, and is updated every month. ERA5 is used by scientists, researchers, and meteorologists to study the atmosphere and climate, and to improve weather forecasting. We have downloaded all 9 predictor variables from ERA5 repository using Google earth engine.

2.1.2 JAXA

The foundation of JAXA GSMaP is the study "Production of a high-precision, high-resolution global precipitation map using satellite data," which was funded by the Japan Science and Technology Agency's Core Research for Evolutional Science and Technology (CREST) (JST). We have downloaded hourly precipitation data for North Indian Region from here.

2.2 Underlying variables

We have used total ten variables for the project from above mentioned sources in order predict cloudburst. The table below shows the correlation of the 9 variables with cloudburst and their units.



S.No	Predicated Variable	Units	Correlation With Cloudburst
1	dewpoint_temperature_2m	K	0.00257314
2	temperature_2m	K	0.002614164
3	surface_latent_heat_flux	J/m2	7.53E-05
4	surface_net_thermal_radiation	J/m2	0.002918215
5	surface_sensible_heat_flux	J/m2	0.000801906
6	u_component_of_wind_10m	m/s	-0.005544649
7	v_component_of_wind_10m	m/s	0.004259487
8	surface_pressure	Pa	0.001181061
9	total_precipitation_hourly	m	0.001713846

Figure 2.1: The correlation of the 9 variables with cloudburst and their units.

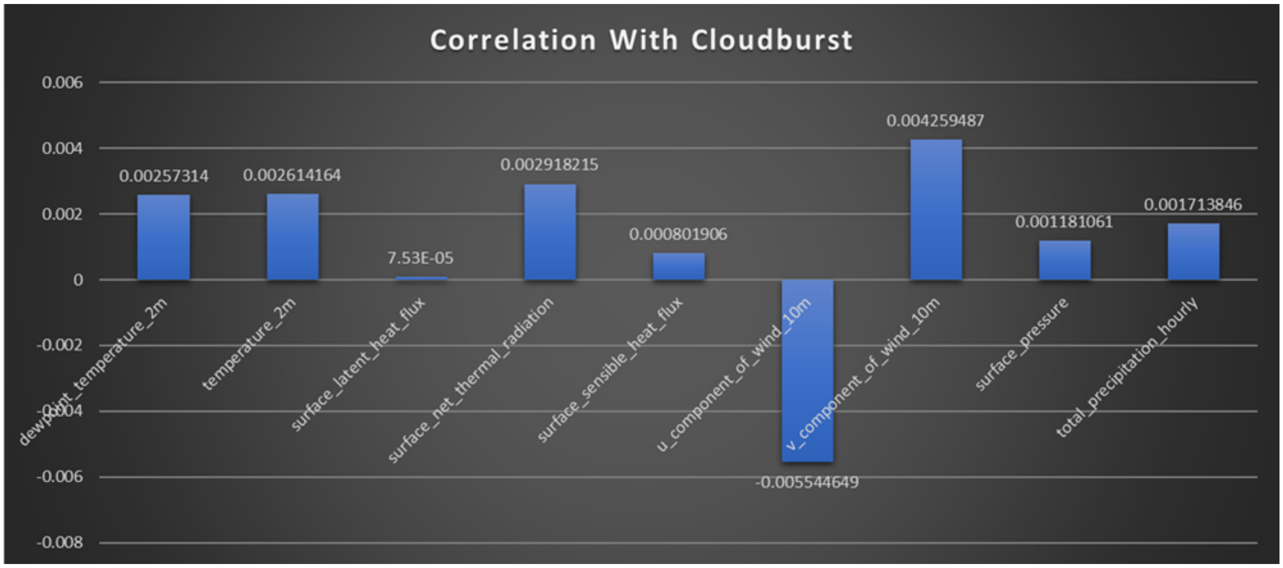


Figure 2.2: Graph of correlation coefficient of predictor variables with cloudburst

2.3 Overview of the data

All the data were downloaded for North Indian region. The grid for the dataset was 28 °N to 37 °N latitude and 72 °E to 81 °E longitude. The data was downloaded for a smaller chunk based on time and then merge it with the help of Dask Array.

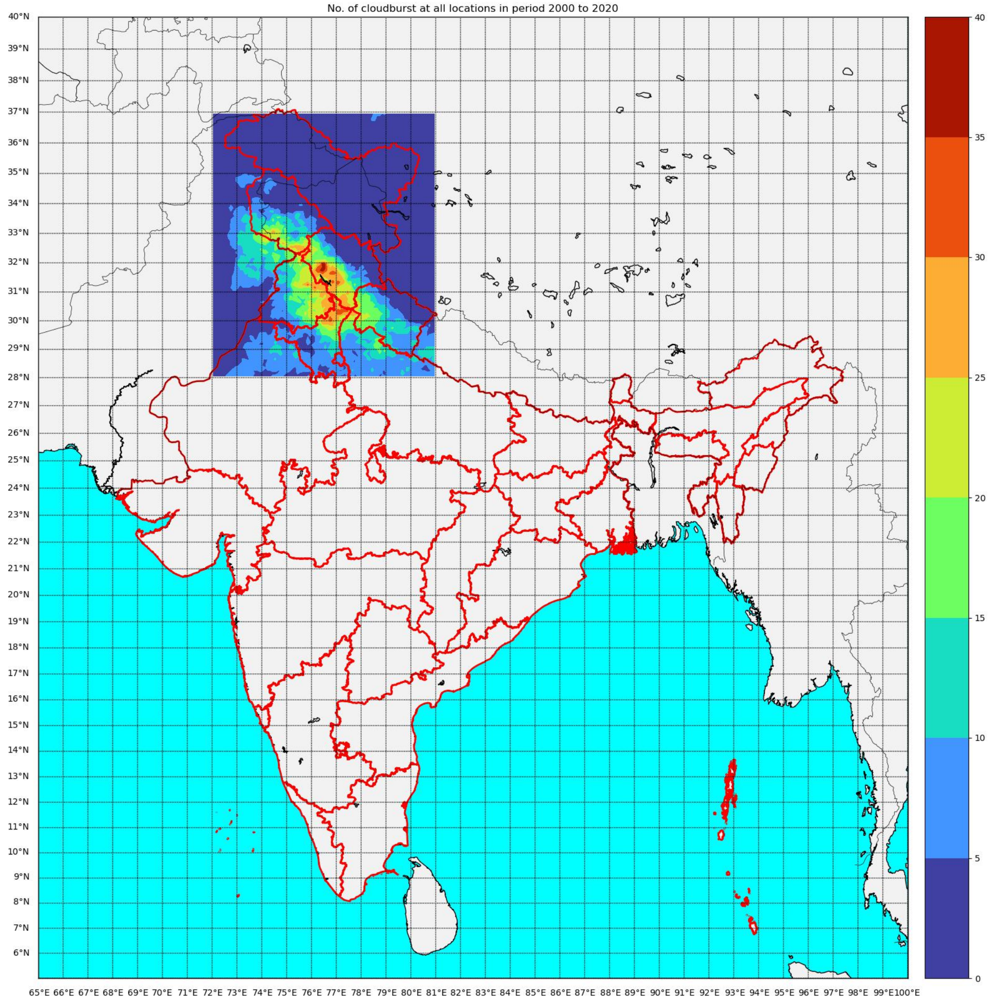

No. of cloudburst at all locations in period 2000 to 2020
40°N
39°N
38°N
37°N
36°N
35°N
34°N
33°N
32°N
31°N
30°N
29°N
28°N
27°N
26°N
25°N
24°N
23°N
22°N
21°N
20°N
19°N
18°N
17°N
16°N
15°N
14°N
13°N
12°N
11°N
10°N
9°N
8°N
7°N
6°N
65°E 66°E 67°E 68°E 69°E 70°E 71°E 72°E 73°E 74°E 75°E 76°E 77°E 78°E 79°E 80°E 81°E 82°E 83°E 84°E 85°E 86°E 87°E 88°E 89°E 90°E 91°E 92°E 93°E 94°E 95°E 96°E 97°E 98°E 99°E 100°E
40
35
30
25
20
15
10
5
0

Figure 2.3: No. of cloudburst at all locations in period 2000 to 2020

2.3.1 Dew Point Temperature (2m)

It is the temperature that the air would need to be cooled to in order for saturation to take place when it is 2 metres above the Earth's surface. It serves as a gauge for atmospheric humidity. It is possible to determine the relative humidity using this information along with temperature and pressure. By interpolating between the lowest model level and the

Earth's surface while taking the atmospheric conditions into consideration, the dew point temperature at 2 metres is computed.

2.3.2 Temperature (2m)

it is the air temperature measured two metres above land, water, or other bodies of surface. By interpolating between the lowest model level and the Earth's surface while taking the atmospheric conditions into consideration, the temperature at 2 metres is estimated.

2.3.3 Surface Latent Heat Flux

It is the latent heat transfer with the surface via turbulent diffusion. From the start of the forecast period through the conclusion of the forecast phase, these variables are added up. Positive downward fluxes are the norm in model theory.

2.3.4 Surface Net Thermal Radiation

It is the surface thermal radiation on a net basis. Field that has accumulated from the start of the forecast time to the conclusion of the forecast step. Positive downward fluxes are the norm for models.

2.3.5 Surface Sensible Heat Flux

It is the heat transfer from the Earth's surface to the atmosphere caused by turbulent air motion (but excluding any heat transfer resulting from condensation or evaporation). The difference in temperature between the surface and the surrounding atmosphere, wind speed, and surface roughness all influence how much sensible heat is transferred. For instance, a sensible heat flow from the land (or ocean) into the atmosphere might result from cold air topping a warm surface. From the start of the forecast time through the conclusion of the forecast step, this single-level variable is added up. joules per square metre are used as the units (J m-2).

2.3.6 U Wind (10m)

The wind's 10-meter component is easterly. At a height of ten metres above the Earth's surface, it is the horizontal speed of air travelling in the direction of the east, measured in metres per second. When comparing this variable to observations, caution should be used because wind data vary on small spatial and temporal scales and are influenced by the local terrain, vegetation, and structures, which are only averaged out in the ECMWF Integrated Forecasting System. The speed and direction of the horizontal 10m wind can be determined by combining this variable with the 10m wind's V component.

2.3.7 V Wind (10m)

The wind's 10m component is to the north. It is the air's horizontal speed, measured in metres per second, as it moves ten metres above Earth's surface in the direction of the north. When comparing this variable to observations, caution should be used because wind data vary on small spatial and temporal scales and are influenced by the local terrain, vegetation, and structures, which are only averaged out in the ECMWF Integrated Forecasting System. The speed and direction of the horizontal 10m wind can be determined by combining this variable with the U component of the 10m wind.

2.3.8 Surface Pressure

It is the force of the atmosphere acting on the surface of land, water, and air (per unit area). The area of the Earth's surface represented at a fixed place is measured by the weight of all the air in a vertical column above it. To determine air density, surface pressure and temperature are frequently combined. Mean sea level pressure, rather than surface pressure, is typically employed for this purpose due to the significant pressure change with height that makes it challenging to observe the low and high pressure systems over hilly regions. This variable's units are Pascals (Pa). Surface pressure is frequently expressed in hPa and sporadically in the outdated millibars, mb (1 hPa = 1 mb = 100 Pa).

2.3.9 Total Precipitation Hourly

Similar to "total precipitation," but only for the specified prediction step and not accumulated. Precipitation is the process through which water vapour in the atmosphere condenses and falls to the ground due to gravity. As a result of these events, it rains and snows.

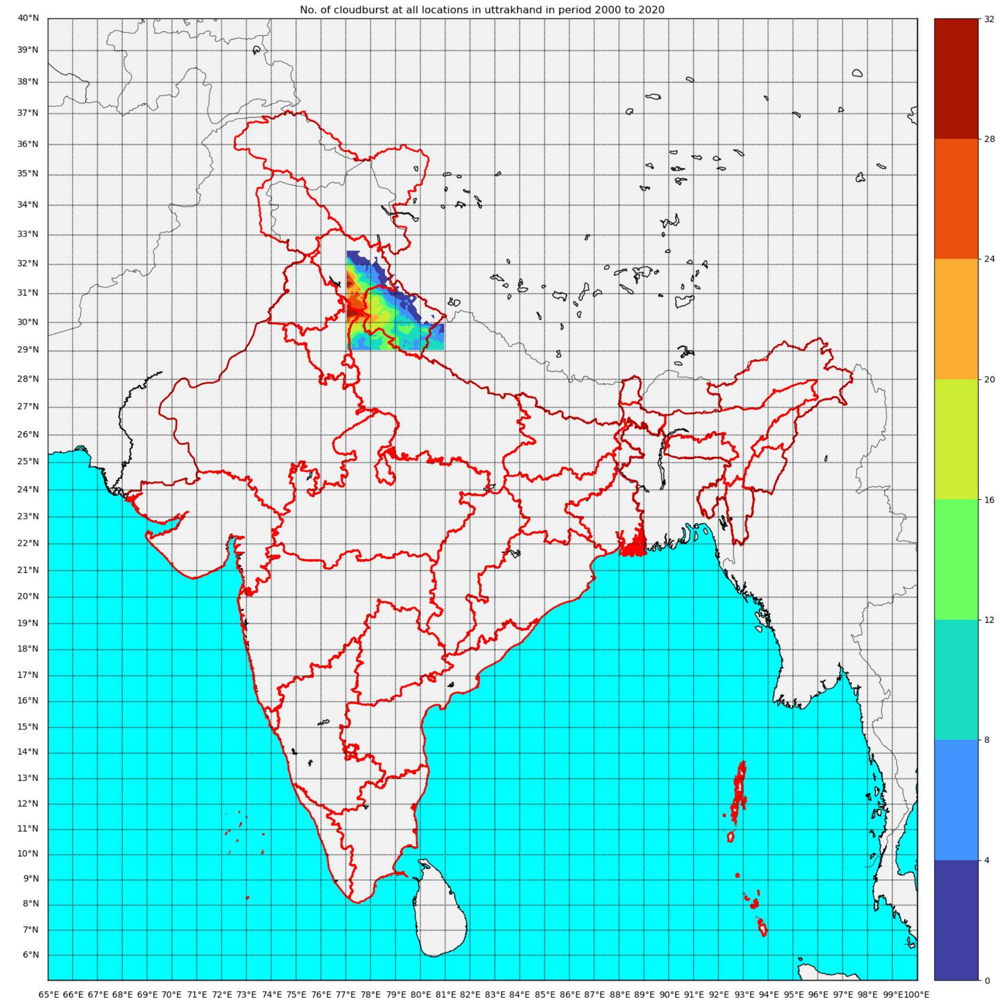

No. of cloudburst at all locations in uttrakhand in period 2000 to 2020
40°N
39°N
38°N
37°N
36°N
35°N
34°N
33°N
32°N
31°N
30°N
29°N
28°N
27°N
26°N
25°N
24°N
23°N
22°N
21°N
20°N
19°N
18°N
17°N
16°N
15°N
14°N
13°N
12°N
11°N
10°N
9°N
8°N
7°N
6°N
65°E 66°E 67°E 68°E 69°E 70°E 71°E 72°E 73°E 74°E 75°E 76°E 77°E 78°E 79°E 80°E 81°E 82°E 83°E 84°E 85°E 86°E 87°E 88°E 89°E 90°E 91°E 92°E 93°E 94°E 95°E 96°E 97°E 98°E 99°E100°E
32
28
24
20
16
12
8
4
0

Figure 2.4: No. of cloudburst at all locations in Uttrakhand in period 2000 to 2020

2.4 Data Preparation & Preprocessing

1. Collected the data from different repositories and merge each variables along the time axis for the whole period of 5 years.

2. Converted all the variables into a fixed resolution ,here it is 0.1°resolution. We have

used xESMF to convert the data.

3. We added one column of cloudburst having values 0 and 1, depending upon the condition that there will be a cloudburst whenever the precipitation will exceed 100mm.

4. We extracted the data of Uttrakhand region from dataset of whole North India and plotted the no. of cloudburst at all locations in Uttrakhand in period 2000 to 2020 fig 2.4 .

5. We found out extracted and the data for the corresponding latitude and longitude for which cloudburst had occured atleast once in past 5 years (2001-2005) in Uttrakhand region.

6. Normalized the data using Min Max Scaler. Equation used -

$$z = \frac{(x - min)}{(max - min)}$$

7. Replaced all the NaN value.

8. We did the re-sampling of data set since it was imbalanced.

9. Split the data into training and testing. We took 75 percent data as training and 25 percent as testing.

Chapter 3

Methodology

3.1 Proposed Model

3.1.1 Model Architecture

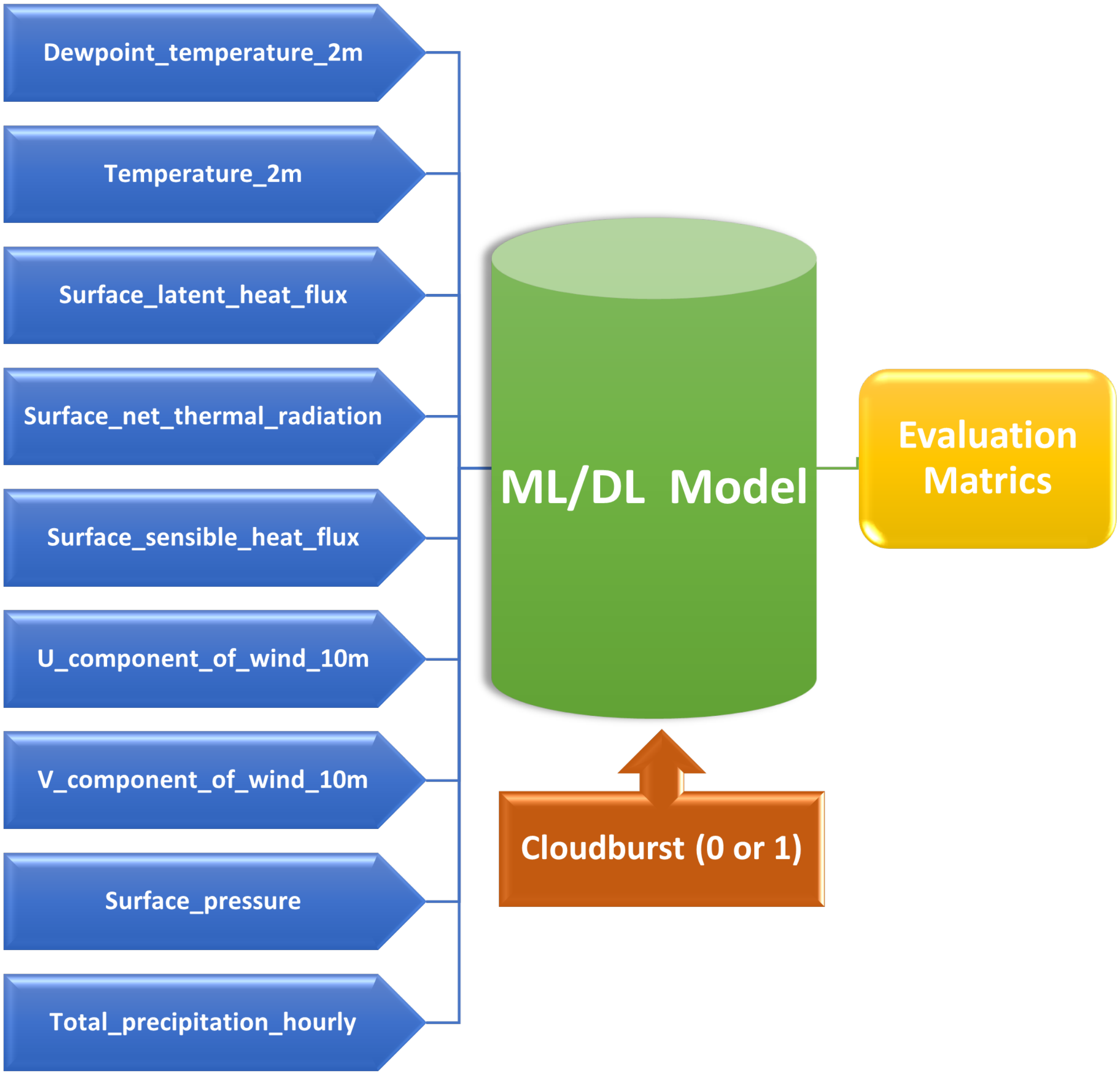


Figure 3.1: Model Architecture for Cloudburst prediction

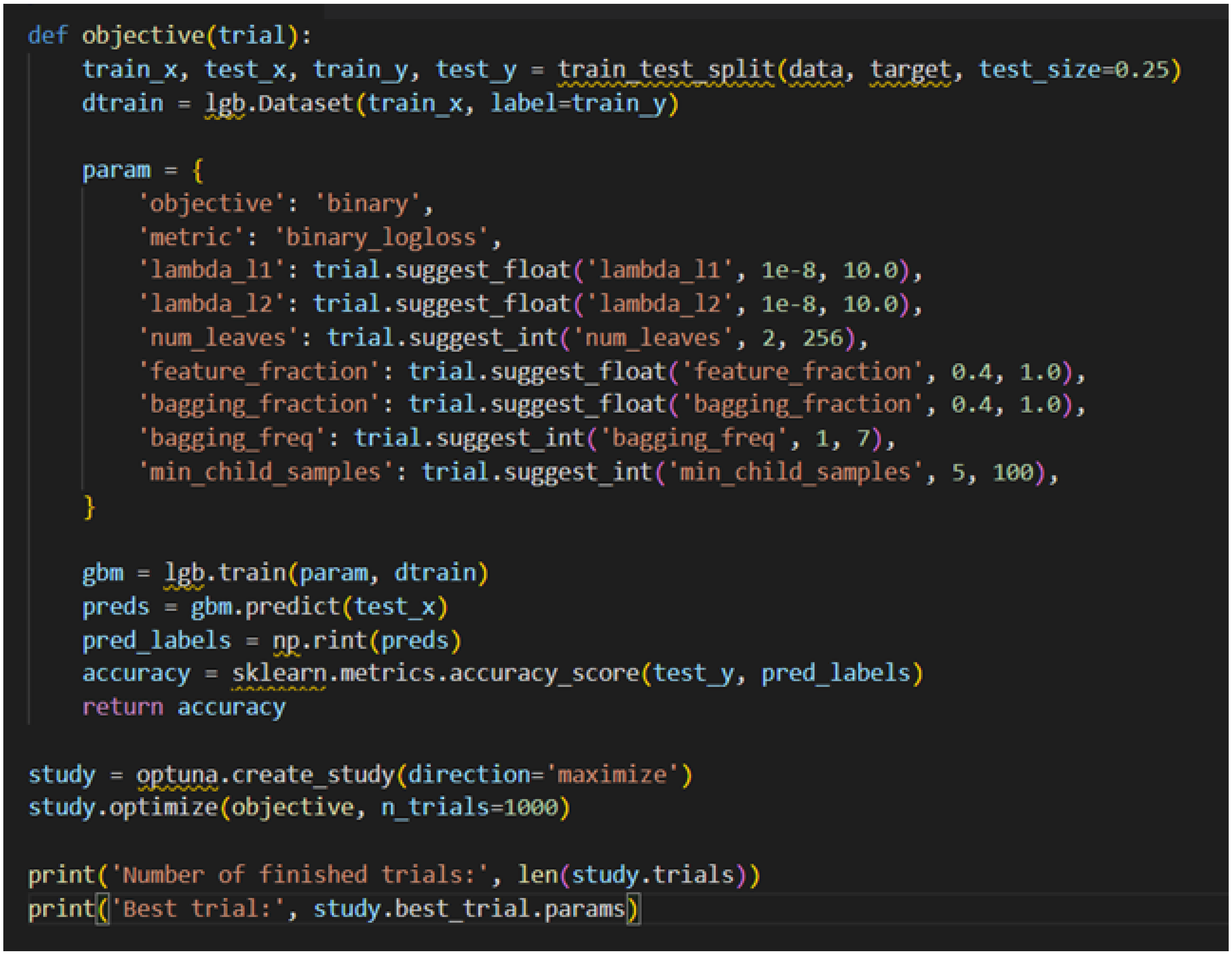
The above mentioned 9 predictor variables are the input variables for the model and cloudburst is the target variable. The Machine Learning or Deep Learning will be trained using input and target variable. For the evaluation part the model is evaluated on test data based on its classification report and AUC-ROC curve, since it is classification task.

3.1.2 Light GBM

A popular, competitive, extremely reliable, and interpretable machine learning approach is gradient-boosted decision trees (GBDT). It performs better than other conventional models in many machine learning applications and is frequently utilised. However, the quantity of samples and features has grown with the introduction of big data. As a result, the GBDT algorithm encountered new difficulties and was unable to deliver outcomes that were effective and accurate enough [6]. Processing massive amounts of data with many features still results in insufficient efficiency and scalability for the GBDT method.The primary issue is that in order to estimate the information gain for every potential split point for each feature, the GBDT method must first scan all data instances. This takes a lot of time. To get around these restrictions, the LightGBM algorithm was devised [6]. This gradient emphasis framework employs a decision tree learning approach based on histograms. It expands upon Gradient-Based One-Sided Sampling (GOSS) and Exclusive Feature Bundles, two more recent methodologies (EFB) [6].GOSS decreases the amount of data instances, eliminates a sizeable portion of data instances with minor gradients, and uses only the remaining data for information gain calculation [6]. However, the number of functions is actually decreased by using an EFB that bundles specific functions. The model is thought to be effective because it can handle vast volumes of data and attain accuracy while training more quickly and using less memory. Additionally, parallel and distributed training are supported. These benefits have led to the widespread use of LightGBM in numerous research fields, and it has produced results that are very encouraging for a variety of machine learning tasks, including image classification, speech recognition [11], social media popularity prediction [5], text classification [7], and online click fraud detection [7].

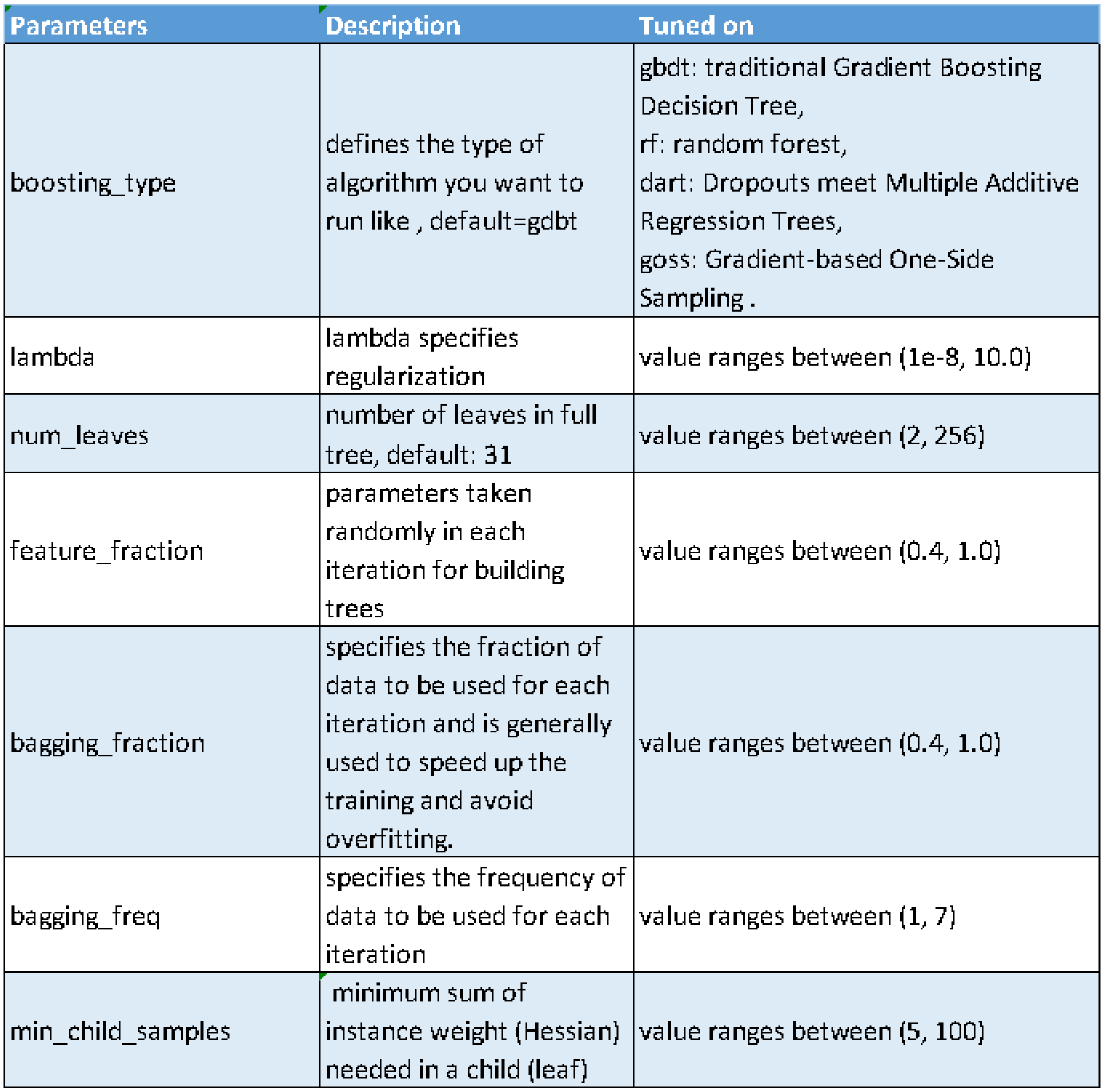
3.1.3 Hyper-parameter Tuning using Optuna

The optimization of hyperparameters is one of the main problems in the development of machine learning models. Performance of the model is directly correlated with proper hyperparameter optimization. An open-source Python package for optimising hyperparameters is called Optuna. Optuna works to balance the sampling and pruning algorithms in the background [1]. For relational parameter sampling, which tries to take advantage of parameter correlation, Optuna employs sample algorithms including the Tree-Structured of Parzen Estimator (TPE), Gaussian Processes (GP), and Covariance Matrix Adaptation (CMA) [1].For the pruning of search spaces, Optuna uses a variation of the Asynchronous Successive Halving (ASHA) algorithm [1].



```python
def objective(trial):
    train_x, test_x, train_y, test_y = train_test_split(data, target, test_size=0.25)
    dtrain = lgb.Dataset(train_x, label=train_y)

    param = {
        'objective': 'binary',
        'metric': 'binary_logloss',
        'lambda_l1': trial.suggest_float('lambda_l1', 1e-8, 10.0),
        'lambda_l2': trial.suggest_float('lambda_l2', 1e-8, 10.0),
        'num_leaves': trial.suggest_int('num_leaves', 2, 256),
        'feature_fraction': trial.suggest_float('feature_fraction', 0.4, 1.0),
        'bagging_fraction': trial.suggest_float('bagging_fraction', 0.4, 1.0),
        'bagging_freq': trial.suggest_int('bagging_freq', 1, 7),
        'min_child_samples': trial.suggest_int('min_child_samples', 5, 100),
    }

    gbm = lgb.train(param, dtrain)
    preds = gbm.predict(test_x)
    pred_labels = np.rint(preds)
    accuracy = sklearn.metrics.accuracy_score(test_y, pred_labels)
    return accuracy

study = optuna.create_study(direction='maximize')
study.optimize(objective, n_trials=1000)

print('Number of finished trials:', len(study.trials))
print('Best trial:', study.best_trial.params)
```

Figure 3.2: Code snippet of Optuna based hyperparameter tuning of our light GBM model



Parameters	Description	Tuned on
boosting_type	defines the type of algorithm you want to run like , default=gdbt	gbdt: traditional Gradient Boosting Decision Tree, rf: random forest, dart: Dropouts meet Multiple Additive Regression Trees, goss: Gradient-based One-Side Sampling .
lambda	lambda specifies regularization	value ranges between (1e-8, 10.0)
num_leaves	number of leaves in full tree, default: 31	value ranges between (2, 256)
feature_fraction	parameters taken randomly in each iteration for building trees	value ranges between (0.4, 1.0)
bagging_fraction	specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting.	value ranges between (0.4, 1.0)
bagging_freq	specifies the frequency of data to be used for each iteration	value ranges between (1, 7)
min_child_samples	minimum sum of instance weight (Hessian) needed in a child (leaf)	value ranges between (5, 100)

Figure 3.3: List of parameters and their description that are being hyper-parameterized in our light GBM model

The table in fig 3.3 shows the list of parameters and their description that are being hyper-parameterized in our light GBM model using Optuna. The remaining parameters of light GBM are taken as default. The algorithm was trained using Apollo High Processing Computer (HPC) of Indian Institute Of Tropical Meteorology,Pune. The specifications of HPC are :- 38,144 Intel Sandy Bridge processors in 2384 compute nodes,149 TB RAM,Infiniband FDR10 interconnect,6 PB of GPFS-based disk storage,1 PB of tape archive capacity.

Chapter 4

Summary and Conclusions

4.1 Results and Conclusions

4.1.1 Results

We observed that the data set was imbalanced, it had 340 occurances of cloudburst in 5 yrs and for its not occurance it was 9903884. So, we did resampling of data.We down-sampled the majority class (0-cloudburst) data and made it equal to minority class (1-cloudburst) as shown in fig 4.1.

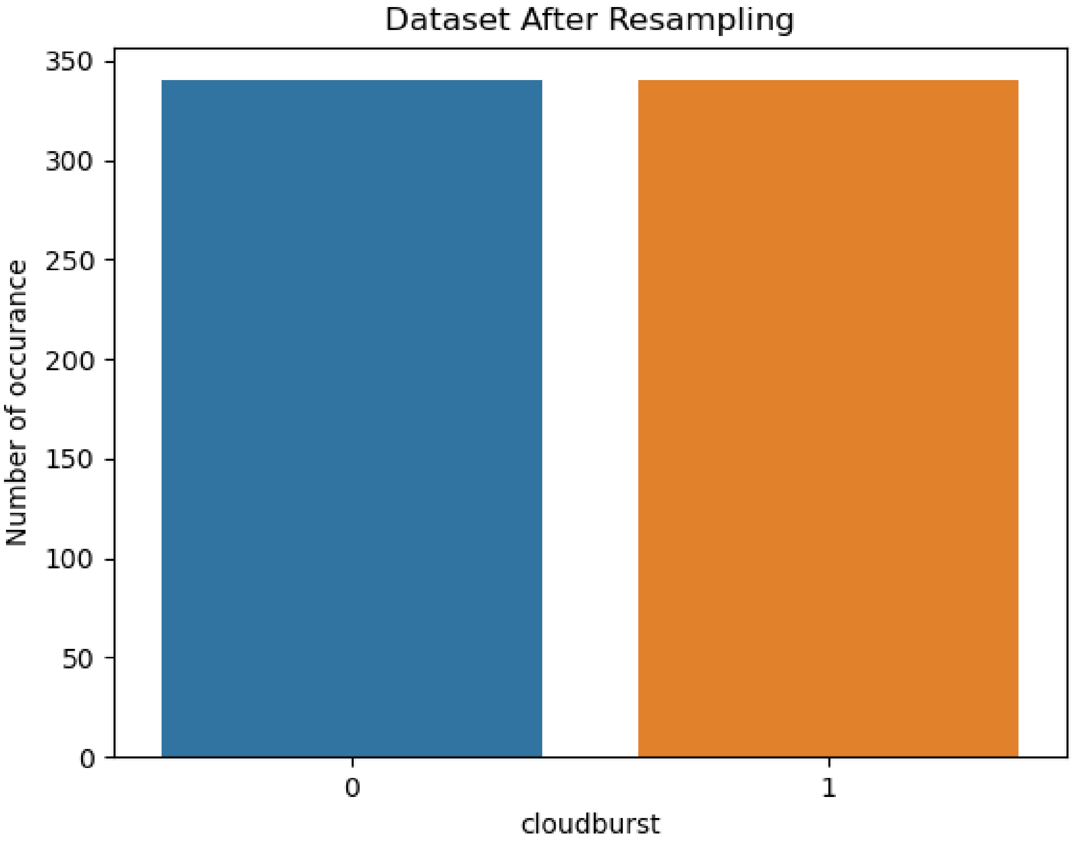


Figure 4.1: Dataset After Resampling

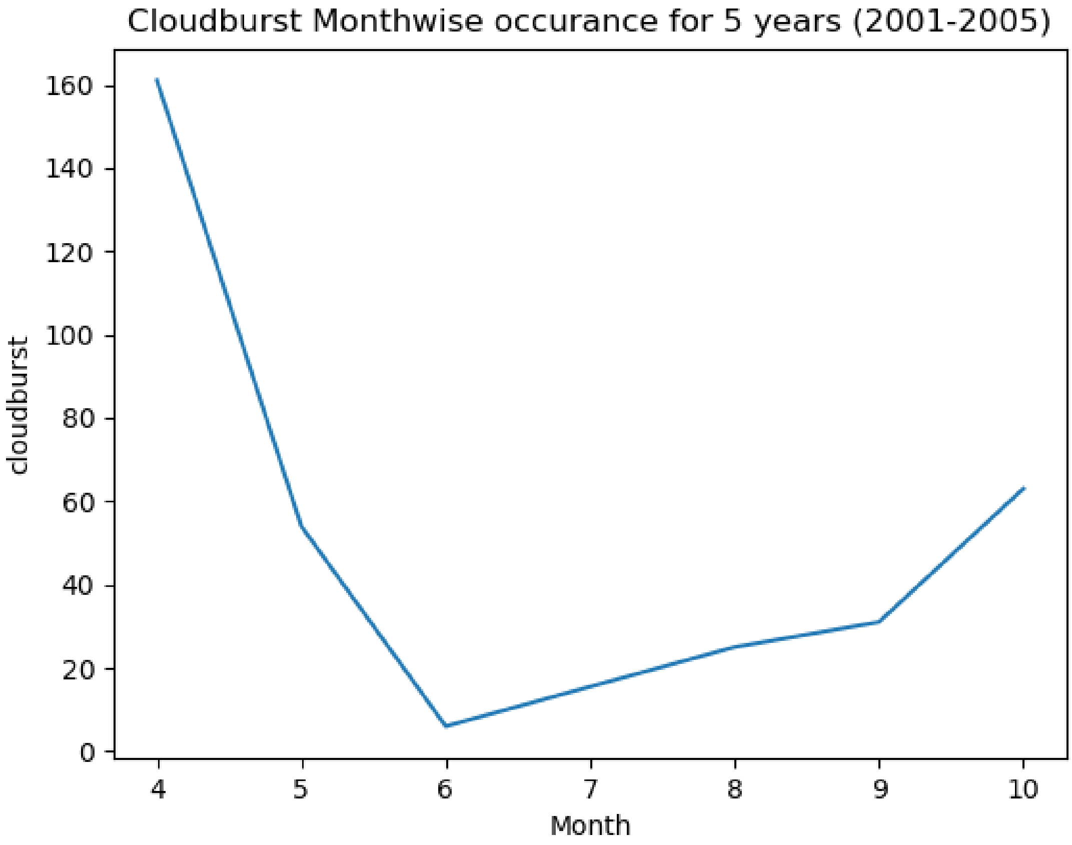
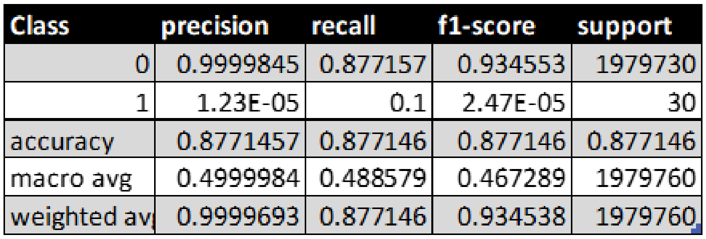


Figure 4.2: Cloudburst monthwise occurance for 5 years (2001-2005)

In fig 4.2 we found the monthly sum of the cloudburst events in duration of 5 years. Then we trained Light GBM Classifier model for 5 years i.e. 2001-2005 and tested it for year 2006.



Class	precision	recall	f1-score	support
0	0.9999845	0.877157	0.934553	1979730
1	1.23E-05	0.1	2.47E-05	30
accuracy	0.8771457	0.877146	0.877146	0.877146
macro avg	0.4999984	0.488579	0.467289	1979760
weighted av	0.9999693	0.877146	0.934538	1979760

Figure 4.3: Classifcation report of Light GBM model

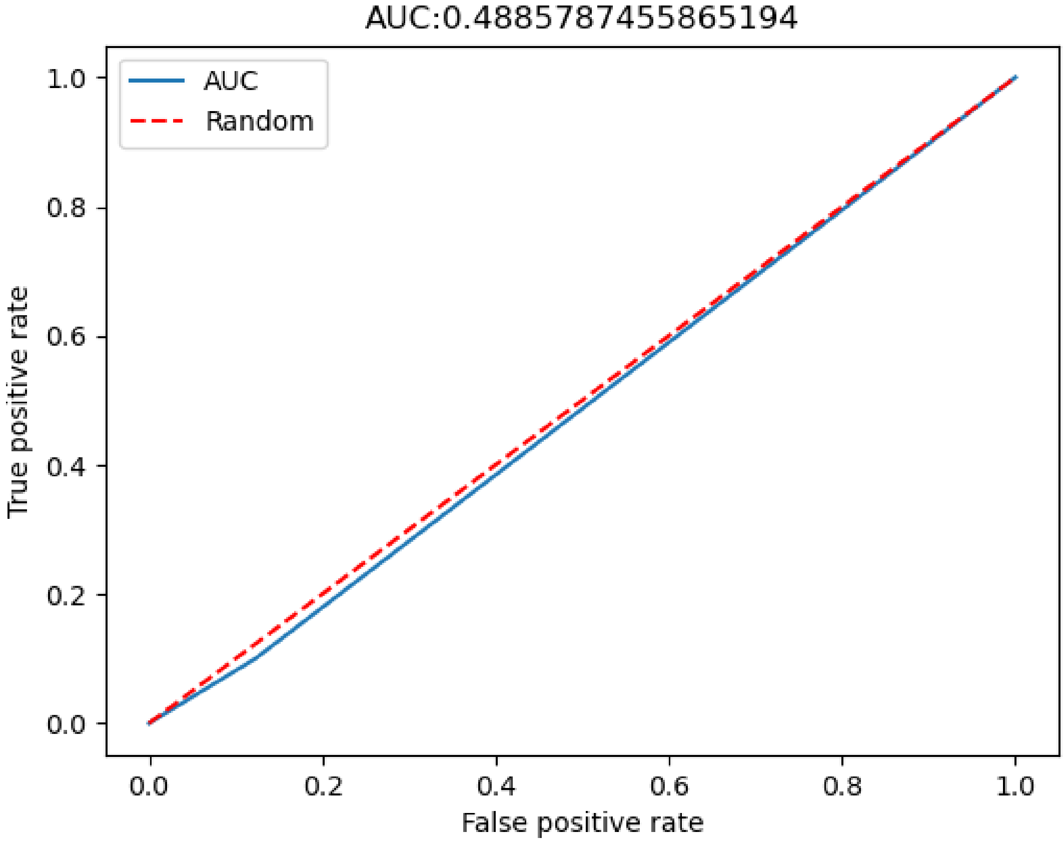


Figure 4.4: AUC curve for Light GBM model

Fig 4.3 and fig 4.4 are the respective classification report and AUC curve for Light GBM model.

4.1.2 Conclusion

From fig 4.3 and fig 4.4 which are classification report and AUC curve for Light GBM model we found our classifier is a weak classifier although its accuracy is 0.877. In 2006 year there were only 30 instances of cloudburst and there were 1979730 instances of not having cloudburst. Our classifier is predicting 0 (no cloudburst) with high precision but not predicting 1 (cloudburst) with same precision. Also its area under curve (AUC) is below 0.5, which needed to be greater than 0.5 for a better classifier. We need to get a better classifier in order to predict with precision.

4.1.3 Future Outlook

- We will try to improve precision of prediction by using more advanced hyper parameter tuned models like XGBoost, CatBoost,TabNet, ResNet,FCTransformer,Artificial Neural Network and much more machine learning regression based models.
- We will try to extract data for more years in order to better train the model.
- We will add some more variables to better predict cloudburst.
- We will be deploying our model once we get the best accuracy.
- We will be using some Ensemble and Encoding techniques as well for better accuracy.

Bibliography

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

[2] Pranab Das. ‘the himalayan tsunami’- cloudburst, flash flood death toll: A geographical postmortem. *IOSR Journal Of Environmental Science, Toxicology And Food Technology*, 7:33–45, 01 2013.

[3] Pranab Das and Das. Global warming, glacial lakes and cloud burst events in garhwal –kumaon himalaya: A hypothetical analysis. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL SCIENCES*, 5, 01 2015.

[4] A.P. Dimri, A. Chevuturi, D. Niyogi, R.J. Thayyen, K. Ray, S.N. Tripathi, A.K. Pandey, and U.C. Mohanty. Cloudbursts in indian himalayas: A review. *Earth-Science Reviews*, 168:1–23, 2017.

[5] Ziliang He, Zijian He, Jiahong Wu, and Zhenguo Yang. Feature construction for posts and users combined with lightgbm for social media popularity prediction. pages 2672–2676, 10 2019.

[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc.

[7] Sushil Khanduri. Cloudbursts over indian sub-continent of uttarakhand himalaya: A traditional habitation input from bansoli, district-chamoli, india. 08 2020.

[8] Kishan Rawat, Smruti Sahu, Sudhir Singh, and Anil Kumar Mishra. Cloudburst analysis in the nainital district, himalayan region, 2021. *Discover Water*, 2, 10 2022.

[9] Bernice Rosenzweig, Benjamin L. Ruddell, Lauren McPhillips, Robert Hobbins, Timon McPhearson, Zhongqi Cheng, Heejun Chang, and Yeowon Kim. Developing knowledge systems for urban resilience to cloudburst rain events. *Environmental Science Policy*, 99:150–159, 2019.

[10] M. Sivagami, P. Radha, and Ananthakrishnan Balasundaram. Sequence model based cloudburst prediction for the indian state of uttarakhand. *Disaster Advances*, 14:1–9, 06 2021.

[11] Jiali Yu, Yuanyuan Qu, Zhongkai Zhang, Qidong Lu, Zhiliang Qin, and Xiaowei Liu. Speech recognition based on concatenated acoustic feature and lightGBM model. In Yi Xie, Jie Tian, Dahong Qian, Yue Lyu, and Kezhi Mao, editors, *Twelfth International Conference on Signal Processing Systems*, volume 11719, page 117190P. International Society for Optics and Photonics, SPIE, 2021.