WEATHER FORECASTING USING DIGITAL IMAGE PROCESSING

A Project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE ENGINEERING

Submitted by

HEMAJA PATOJU (316126510102)

G.SAI CHARAN (316126510075)

M.KALI CHARAN (31612651010094)

PRANITA JAGTAP (316126510104)

Under the guidance of

Mrs G.Gowripushpa

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES
(UGC AUTONOMOUS)
(*Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade*)
Sangivalasa, bheemili mandal, visakhapatnam dist.(A.P)

2019-2020

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES
(UGC AUTONOMOUS)
(*Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade*)
Sangivalasa, bheemili mandal, visakhapatnam dist.(A.P)

BONAFIDE CERTIFICATE

This is to certify that the project report entitled "**WEATHER FORECASTING USING DIGITAL IMAGE PROCESSING"** submitted by **Hemaja Patoju (316126510102), G.Sai Charan (316126510075), M.Kali Charan (316126510094), Pranita Jagtap (316126510104)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science Engineering** of Anil Neerukonda Institute of technology and sciences (A), Visakhapatnam is a record of bonafide work carried out under my guidance and supervision.

Project Guide

Mrs. G. Gowripushpa
Assistant Professor
Department of CSE
ANITS

Head of the Department

Dr. R.Sivaranjani
Professor
Department of CSE
ANITS

DECLARATION

We, **Hemaja Patoju, G.Sai Charan, M.Kali Charan, Pranita Jagtap**, of final semester B.Tech., in the department of Computer Science and Engineering from ANITS, Visakhapatnam, hereby declare that the project work entitled **WEATHER FORECASTING USING DIGITAL IMAGE PROCESSING** is carried out by us and submitted in partial fulfillment of the requirements for the award of **Bachelor of Technology in Computer Science Engineering** , under Anil Neerukonda Institute of Technology & Sciences(A) during the academic year 2016-2020 and has not been submitted to any other university for the award of any kind of degree.

HEMAJA PATOJU	316126510102
G.SAI CHARAN	316126510075
M.KALI CHARAN	316126510094
PRANITA JAGTAP	316126510104

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement always boosted the morale. We take a great pleasure in presenting a project, which is the result of a studied blend of both research and knowledge.

We first take the privilege to thank, **Dr. R.Sivaranjani**, Head of our Department, for permitting us in laying the first stone of success and providing the lab facilities, we would also like to thank the other staff in our department and lab assistants who directly or indirectly helped us in successful completion of the project.

We feel great to thank **Mrs G. Gowripushpa** , who are our project guides and who shared their valuable knowledge with us and made us understand the real essence of the topic and created interest in us to work day and night for the project; we also thank our B.Tech coordinator **Dr.K.S.Deepthi**, for her support and encouragement.

We also would like to thank **B.Udaya Lakshmi,** who extended their part of support in successful completion of the project.

PROJECT STUDENTS

Hemaja Patoju(316126510102)
G.Sai Charan(316126510075)
M.Kali Charan(316126510094)
Pranita Jagtap(316126510104)

ABSTRACT

To predict the conditions of the atmosphere for a given location Weather Forecasting is used. It is the application of science and technology. Weather forecast is more helpful for people as it predicts how the future weather is going to be and people may plan accordingly. Farmers will be the most beneficial one's as they may know the rainfall prediction and grow crops accordingly. The weather forecast can be done in many ways like using the previous data or analyzing the current clouds. The authors predict the weather using the status of the clouds. The author used methodologies like Normalization, Clustering, and Cloud mask algorithm to predict the weather more accurately. Normalization is done using RGB values of each pixel. In many fields of research and in industrial and military applications Digital-image processing has become economical.

Keywords

Weather forecasting, normalization, clustering, cloud mask algorithm.

CONTENTS

CHAPTER 5 CONCLUSION AND FUTURE WORK

LIST OF FIGURES

LIST OF ABBREVATIONS

MSS	Multispectural Scanner System
JPEG	Joint Photographic Experts Group
JPG	Joint Photographic Experts Group
PNG	Portable Network Graphics
LISS	Low- Intensity Steady-Rate
NARX	Nonlinear Autoregressive models with exogenous input

1. INTRODUCTION

1.1 INTRODUCTION

Weather forecasting means predicting the weather and telling how the weather changes with change in time. Change in weather occurs due to movement or transfer of energy. Many meteorological patterns and features like anticyclones, depressions, thunderstorms, hurricanes and tornadoes occur due to the physical transfer of heat and moisture by convective processes. Clouds are formed by evaporation of water vapor. As the water cycle keeps on evolving the water content in the clouds increases which in turn leads to precipitation. This is how the convective process happens and also the change in weather. Many factors like temperature, rainfall, pressure, humidity, sunshine, wind and cloudiness are considered for predicting the weather. It is also possible to identify the different types of clouds associated with different patterns of weather. These patterns of weather help in predicting the weather forecast.

In the past, people used barometric pressure, current weather conditions, sky condition to predict whereas now there are many computer based models that consider the atmospheric factors to predict the weather. These methods are not accurate and the reason is due to the chaotic nature of the atmosphere as it keeps on changing. Even predicting weather for a longer period of time will not be accurate that is why most of the current forecasting[1] models predict weather only for a couple of days not more than 10. The accuracy gets reduced with increase in time.

Weather forecasting isn't a purely mechanical linear process, that standard practices and procedures will be directly applied. Forecaster's job is predicated on theoretical background and lab work which needs several years of study but mainly day-to-day practice inside a weather forecasting service having a particular technical environment. The work of the forecasters has evolved significantly over the years to require advantage of both scientific and technological improvements. The skill of numerical models has improved such a

lot that some centers are automating routine forecasts to permit forecasters to specialize in high impact weather or areas where they can add significant value. So it's dangerous to see a regular thanks to achieve weather forecasts.

1.1.1 DIGITAL IMAGE PROCESSING

The term digital image processing generally refers to processing of a two-dimensional picture by a digital computer. In a broader context, it implies digital processing of any two-dimensional data. A digital image is an array of real numbers represented by a finite number of bits. The principle advantage of Digital Image Processing methods is its versatility, repeatability and the preservation of original data precision.

Pixel:

Pixel is the smallest element of an image. Each pixel corresponds to any one value. In an 8-bit gray scale image, the value of the pixel between 0 and 255. The values of a pixel at any point correspond to the intensity of the light photons striking at that point. Each pixel stores a value proportional to the light intensity at that particular location.

Digital image:

A digital image is nothing more than data— indicating variations of red, green, and blue at a particular location on a grid of pixels.

Gray level:

The value of the pixel at any point denotes the intensity of image at that location, and that is also known as gray level.Generally to convert an image to gray scale[3], the equation that was used previously is :

Grayscale = (Red + Green + Blue / 3).

But as red has more wavelength we use the equation:

Grayscale = ((0.3 * R) + (0.59 * G) + (0.11 * B)).

Image Processing Techniques:

The basic definition of image processing refers to processing of digital image, i.e removing the noise and any kind of irregularities present in an image using the digital computer. The noise or irregularity may creep into the image either during its formation or during transformation etc. For mathematical analysis, an image may be defined as a two dimensional function f(x,y) where x and y are spatial (plane) coordinates, and the amplitude of f at any pair of coordinates (x, y) is called the intensity or gray level of the image at that point. When x, y, and the intensity values of f are all finite, discrete quantities, we call the image a digital image. It is very important that a digital image is composed of a finite number of elements, each of which has a particular location and value. These elements are called picture elements, image elements, pels, and pixels. Pixel is the most widely used term to denote the elements of a digital image.

Various techniques have been developed in Image Processing during the last four to five decades. Most of the techniques are developed for enhancing images obtained from unmanned space crafts, space probes and military reconnaissance flights. Image Processing systems are becoming popular due to easy availability of powerful personnel computers, large size memory devices, graphics software etc.

The various Image Processing techniques are:

1) Image preprocessing
2) Image enhancement
3) Image segmentation
4) Feature extraction
5) Image classification
6) Image compression

7) Image restoration
8) Image acquisition
9) Image representation
10) Image fusion
11) Linear filtering

1.1.1.1. IMAGE PREPROCESSING:

In image preprocessing, image data recorded by sensors on a satellite restrain errors related to geometry and brightness values of the pixels. These errors are corrected using appropriate mathematical models which are either definite or statistical models. Image enhancement is the modification of image by changing the pixel brightness values to improve its visual impact. Image enhancement involves a collection of techniques that are used to improve the visual appearance of an image, or to convert the image to a form which is better suited for human or machine interpretation.

Sometimes images obtained from satellites and conventional and digital cameras lack in contrast and brightness because of the limitations of imaging sub systems and illumination conditions while capturing image. Images may have different types of noise. In image enhancement, the goal is to accentuate certain image features for subsequent analysis or for image display.

Examples include contrast and edge enhancement, pseudo-coloring, noise filtering, sharpening, and magnifying. Image enhancement is useful in feature extraction, image analysis and an image display. The enhancement process itself does not increase the inherent information content in the data. It simply emphasizes certain specified image characteristics. Enhancement algorithms are generally interactive and application dependent.

Some of the enhancement techniques are:

a. Contrast Stretching

b. Noise Filtering

c. Histogram modification

a. Contrast Stretching

Some images (eg. Over water bodies, deserts, dense forests, snow, clouds and under hazy conditions over heterogeneous regions) are homogeneous i.e., they do not have much change in their levels. In terms of histogram representation, they are characterized as the occurrence of very narrow peaks. The homogeneity can also be due to the incorrect illumination of the scene . Ultimately the images hence obtained are not easily interpretable due to poor human perceptibility. This is because there exists only a narrow range of gray-levels in the image having provision for wider range of gray-levels. The contrast stretching methods are designed exclusively for frequently encountered situations. Different stretching techniques have been developed to stretch the narrow range to the whole of the available dynamic range.

b. Noise Filtering

Noise Filtering is used to filter the unnecessary information from an image. It is also used to remove various types of noises from the images. Mostly this feature is interactive. Various filters like low pass, high pass, mean, median etc., are available .

c. Histogram Modification

Histogram has a lot of importance in image enhancement. It reflects the characteristics of image. By modifying the histogram, image characteristics can be modified. One such example is Histogram

Equalization. Histogram equalization is a nonlinear stretch that redistributes pixel values so that there is approximately the same number of pixels with each value within a range. The result approximates a flat histogram. Therefore, contrast is increased at the peaks and lessened at the tails.

1.1.1.2 IMAGE ENHANCEMENT

The aim of image enhancement is to improve the interpretability or perception of information in images for human viewers, or to provide `better' input for other automated image processing techniques.

Image enhancement is the desired improvement of image quality. Physiological experiments have shown that very small changes in luminance are recognized by the human visual system in regions of continuous grey-tones and are not seen at all in regions of some discontinuities. Therefore, a design goal for image enhancement is often to smooth images in more uniform regions but to preserve edges.

To better the information substance of the image, the visible effect which the image may be having on the interpretation with an object is altered. Some of the routines used to enhance the images are:

(1) Contrast enhancement – A simple linear transformation, called contrast stretch, is used to enhance the contrast of a displayed image by expanding the original grey level range.

(2) Spatial filtering – To enhance naturally occurring features such as fractures, faults, joints.

(3) Density slicing --Continuous grey tone range is converted into a sequence of density ranges, individually conforming to a particular digital interval. Each slice is given an individual color.

(4) False color composite images --Of three bands (MSS bands 4, 5, and 7), increase the amount of information available for interpretation.

1.1.1.3 IMAGE SEGEMENTATION

Segmentation is one of the key problems in image processing. Image segmentation is the process that subdivides an image into its constituent parts or objects. The level to which this subdivision is carried out depends on the problem being solved, i.e., the segmentation[2] should stop when the objects of interest in an application have been isolated e.g., in autonomous air-to-ground target acquisition, suppose our interest lies in identifying vehicles on a road, the first step is to segment the road from the image and then to segment the contents of the road down to potential vehicles. Image thresholding techniques are used for image segmentation.

After thresholding a binary image is formed where all object pixels have one gray level and all background pixels have another – generally the object pixels are 'black' and the background is 'white'. The best threshold is the one that selects all the object pixels and maps them to 'black'. Various approaches for the automatic selection of the threshold have been proposed.

Thresholding can be defined as mapping of the gray scale into the binary set {0, 1} :

$$S(x, y) = \begin{cases} 0, & \text{if } g(x, y) < T(x, y) \\ 1, & \text{if } g(x, y) >= T(x, y) \end{cases}$$

where S(x, y) is the value of the segmented image,

g(x, y) is the gray level of the pixel (x, y) and

T(x, y) is threshold value at the coordinates (x, y).

In the simplest case T(x, y) is coordinate independent and a constant for the whole image. It can be selected, for instance, on the basis of the gray level histogram. When the histogram has two pronounced maxima, which reflect gray levels of object(s) and background, it is possible to select a single threshold for the entire image. A method which is based on this idea and uses a correlation criterion to select the best threshold, is described below. Sometimes gray level histograms have only one maximum. This can be caused, e.g., by inhomogeneous illumination of various regions of the image. In such case it is impossible to select a single thresholding value for the entire image and a local binarization technique must be applied. General methods to solve the problem of binarization of in homogeneously illuminated images, however, are not available.

Segmentation of images involves sometimes not only the discrimination between objects and the background, but also separation between different regions. One method for such separation is known as watershed segmentation.

1.1.1.4 FEATURE EXTRACTION

The feature extraction[3] techniques are developed to extract features in synthetic aperture radar images. This technique extracts high-level features needed in order to perform classification of targets. Features are those items which uniquely describe a target, such as size, shape, composition, location etc. Segmentation techniques are used to isolate the desired object from the scene so that measurements can be made on it subsequently. Quantitative measurements of object features allow classification and description of the image.

When the pre-processing and the desired level of segmentation has been achieved, some feature extraction technique is applied to the segments to obtain features, which is followed by application of classification and post processing techniques. It is essential to focus on the feature extraction phase as it has an observable impact on the efficiency of the recognition system. Feature selection of a feature extraction method is the single most important factor in achieving high recognition performance.

Feature extraction has been given as "extracting from the raw data information that is most suitable for classification purposes, while minimizing[4] the within class pattern variability and enhancing the between class pattern variability". Thus, selection of a suitable feature extraction technique according to the input to be applied needs to be done with utmost care. Taking into consideration all these factors, it becomes essential to look at the various available techniques for feature extraction in a given domain, covering vast possibilities of cases.

1.1.1.5 IMAGE CLASSIFICATION

The simulation results showed that the proposed algorithm performs better with the total transmission energy metric than the maximum number of hops metric. The proposed algorithm provides energy efficient path for data transmission and maximizes the lifetime of entire network. As the performance of the proposed algorithm is analyzed between two metrics in future with some modifications in design considerations the performance of the proposed algorithm can be compared with other energy efficient algorithm. We have used very small network of 5 nodes, as number of nodes increases the complexity will increase. We can increase the number of nodes and analyze the performance.

Image classification is the labeling of a pixel or a group of pixels based on its grey value. Classification is one of the most often used methods of information extraction. In Classification, usually multiple features are used for a set of pixels i.e., many images of a particular object are needed. In Remote Sensing area, this procedure assumes that the imagery of a specific geographic area is collected in multiple regions of the electromagnetic spectrum and is in good registration. Most of the information extraction techniques rely on analysis of the spectral reflectance properties of such imagery and employ special algorithms designed to perform various types of 'spectral analysis'. The process of multispectral classification can be performed using either of the two methods: Supervised or Unsupervised .

In Supervised classification[4], the identity and location of some of the land cover types such as urban, wetland, forest etc., are known as priori through a combination of field works and toposheets. The analyst attempts to locate specific sites in the remotely sensed data that represents homogeneous examples of these land cover types. These areas are commonly referred as TRAINING SITES because the spectral characteristics of these known areas are used to 'train' the classification algorithm for eventual land cover mapping of reminder of the image. Multivariate statistical parameters are calculated for each training site. Every pixel both within and outside these training sites is then evaluated and assigned to a class of which it has the highest likelihood of being a member.

In an Unsupervised classification, the identities of land cover types has to be specified as classes within a scene are not generally known as priori because ground truth is lacking or surface features within the scene are not well defined. The computer is required to group pixel data into different spectral classes according to some statistically determined criteria.

The comparison in medical area is the labeling of cells based on their shape, size, color and texture, which act as features. This method is also useful for MRI images.

1.1.1.6 IMAGE COMPRESSION

Image compression signifies compression of the records among the digital images. Image compression eliminates duplication of the data so that it will be stored and transmitted in an effective way. Image compression might be lossy and lossless. In lossless compression before and after compression the quality of data remains consistent. In lossy compression the quality of data decreases after applying the compression techniques. Lossless compression is mostly used for medical imaging, technical drawing contents and for archival purposes etc. Lossy approaches are used in those environments in which minor loss of quality is acceptable to accomplish a considerable reduction in bit rate. The most widespread technique for compression is JPEG which compresses full color or gray scale images.

This method divides the image into eight by eight blocks. These blocks are divided in such a way so that no overlapping is formed among them. JPEG use discrete cosine transforms technique for compression. There is another technique for compression known as Wavelet transform. Through wavelet data is 44 divided into different frequency components and then further study is done for each component. Wavelets have advantages over traditional Fourier approaches in examining physical circumstances.

1.1.1.7 IMAGE RESTORATION

Image restoration is a method through which a corrupted and noisy image is processed in such a way that a perfect image is constructed.

Thus, restoration[5] rebuilds those images whose quality is despoiled due to noise or system error. There are various causes for degradation such as noise from the sensor, camera misfocus and atmospheric disturbance. There are two types of procedures are used to restore the image. One technique is to model the picture whose quality is degraded via some reasons. Another technique known as image enhancement, it increases the quality of image by applying various filters.

Prior knowledge of degradation is necessary to restore the image. The following figure showing the degradation and restoration activity. Restoration of the images might be achieved via two types of model namely degradation Model and restoration model. In the following diagram f(x,y) is the original image which is degraded by some activities. After this on the degraded image various functions are applied in order to restore the image. Figure 4: Degradation- restoration model.

1.1.1.8 IMAGE ACQUISITION

The first phase of every visualization scheme is the image acquisition phase. When the image is obtained then various processes are applied on the image. Basically, an image acquisition is a process through which images are retrieved from various resources. The most common method for image acquisition is real time acquisition method. This method creates a pool of files which are processed automatically. An image acquisition method creates 3D geometric data.

1.1.1.9 IMAGE REPRESENTATION

Image representation means converting the raw data in such a way so that computer processing can apply on it. Basically, two types of techniques are used to represent the pictures. Boundary representation and region representation. Boundary representation display the internal shape

of the picture. It means the main concern of boundary representation method is to display what is the shape of the object, whether it is corner, rounded or any other shape. Region representation is used when the main concern is about the internal properties.

Depends upon level of processing of images via machine there are four methods of image representation such as pixel based, Block based, Region based and Hierarchical based. Image representation is appropriate for the formation of entities, knowledge based models which must be extracted from image databases that are created using predefined decision rules.

1.1.1.10 IMAGE FUSION TECHNIQUES

The satellites cover different portions of the electromagnetic spectrum and record the incoming radiations at different spatial, temporal, and spectral resolutions. Most of these sensors operate in two modes: multispectral mode and the panchromatic mode. The panchromatic mode corresponds to the observation over a broad spectral band (similar to a typical black and white photograph) and the multispectral (color) mode corresponds to the observation in a number of relatively narrower bands. For example in the IRS – 1D, LISS III operates in the multispectral mode. It records energy in the green (0.52 – 0.59 μm), red (0.62-0.68 μm), near infrared (0.77- 0.86 μm) and mid-infrared (1.55 – 1.70 μm). In the same satellite PAN operates in the panchromatic mode. SPOT is another satellite, which has a combination of sensor operating in the multispectral and panchromatic mode.

Above information is also expressed by saying that the multispectral mode has a better spectral resolution than the panchromatic mode. Now coming to the spatial resolution, most of the satellites are such that the panchromatic mode has a better spatial resolution than the multispectral mode, for e.g. in IRS -1C, PAN has a spatial resolution of

5.8 m whereas in the case of LISS it is 23.5 m. Better is the spatial resolution, more detailed information about a landuse is present in the imagery, hence usually PAN data is used for observing and separating various feature. Both these type of sensors have their particular utility as per the need of user. If the need of the user is to separate two different kinds of landuses, LISS III is used, whereas for a detailed map preparation of any area, PAN imagery is extremely useful. Image Fusion is the combination of two or more different images to form a new image (by using a certain algorithm).

Commonly applied Image Fusion Techniques are :

1. IHS Transformation
2. PCA
3. Brovey Transform
4. Band Substitution

1.1.1.11 LINEAR FILTERING

Linear filtering is one of the most powerful image enhancement methods. It is a process in which part of the signal frequency spectrum is modified by the transfer function of the filter. In general, the filters under consideration are linear and shift-invariant, and thus, the output images are characterized by the convolution sum between the input image and the filter impulse response; that is:

$$y(m, n) = \Sigma i = 0M\Sigma j = 0Nh(m-i, n-j)x(i, j) = h(m, n)** x(m, n),$$

where the following is true:

The y(m, n) is the output image
The h(m, n) is the filter impulse response.

The x(m, n) is the input image.

For example, low-pass filtering has the effect of smoothing an image. On the other hand, high-pass filtering usually sharpens the edges of an image. They can even be used for edge detection, which is used in image analysis algorithms.

The image filtering can be carried out either in the spatial domain, as above equation, or in the frequency domain, using the discrete Fourier transform (DFT) (Mersereau and Dudgeon, 1984; Oppenheim and Schaffer, 1989). For filtering using the DFT, we use the well known property that the DFT of the circular convolution of two sequences is equal to the product of the DFTs of the two sequences. That is, for $y(m,n)$ defined as above equation , provided that a DFT of sufficient size is used, we have that:

$$DFT\{y(m, n)\} = DFT\{h(m, n)\}\ DFT\ \{x(m,n)\}.$$

Therefore, one can perform image filtering in the frequency domain by modifying conveniently the DFT of the image and taking the inverse transformation.

1.1.2. MACHINE LEARNING

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning[10] is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application

domains to the field of machine learning. Data mining is related field of study, focusing the exploratory data analysis using unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

MACHINE LEARNING TECHNIQUES:

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. For early tasks that humans assigned to computers, it was possible to create algorithms telling the machine how to execute all needed steps to solve the problem in hand. So on the computer's part, no learning was needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than have human programmers specify every needed step.

Early classifications for machine learning approaches sometimes divided them into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system. These were:

1) Supervised Learning
2) Unsupervised Learning
3) Semi Supervised Learning
4) Reinforcement Learning

1.1.2.1. SUPERVISED LEARNING:

Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is

commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim

Thinking of supervised learning with the concept of function approximation, where basically we train an algorithm and in the end of the process we pick the function that best describes the input data, the one that for a given X makes the best estimation of y (X -> y).
Most of the time we are not able to figure out the true function that always make the correct predictions and other reason is that the algorithm rely upon an assumption made by humans about how the computer should learn and this assumptions introduce a bias.

The human experts acts as the teacher where we feed the computer with training data containing the input/predictors and we show it the correct answers (output) and from the data the computer should be able to learn the patterns.

Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features such that we can predict the output values for new data based on those relationships which it learned from the previous data sets.
Some common algorithms are:

1) Linear Regression
2) Logistic Regression
3) Decision Tree
4) Random Forest
5) KNN
6) SVM
7) Naïve Bayes

1.1.2.1.1 Linear Regression

To understand the working functionality of this algorithm, imagine how you would arrange random logs of wood in increasing order of their weight. There is a catch; however – you cannot weigh each log. You have to guess its weight just by looking at the height and girth of the log (visual analysis) and arrange them using a combination of these visible parameters. This is what linear regression is like.

In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and represented by a linear equation Y= a *X + b.

Where Y is the Dependent Variable

a is the Slope

X is the Independent Variable

b is the Intercept

The coefficients a & b are derived by minimizing the sum of the squared difference of distance between data points and the regression line.

Real life applications of Linear Regrerssion:

1) Risk Management in financial services or insurance domain
2) Predictive Analytics
3) Econometric
4) Epidemiology
5) Weather data analysis
6) Customer survey results analysis

1.1.2.1.2 Logestic Regression

Logistic Regression[11] is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logistic function. It is also

called logistic regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable.

The sigmoid function, also called the logistic function, gives an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1.

o- (x) = 1/ 1+epow(-x)

If the curve goes to positive infinity, y predicted will become 1.

If the curve goes to negative infinity, y predicted will become 0.

If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it like 0 or NO.
If the output is 0.75, we can say in terms of probability as: There is a 75 percent chance that patient will suffer from cancer.

Real life applications of Logistic Regression:

1) Cancer Detection

2) Trauma and Injury Severity Score

3) Image Segmentation and Categorization

4) Geographic Image Processing

5) Handwriting recognition

6) Prediction whether a person is depressed based on bag of words from the corpus.

1.1.2.1.3 DECISION TREE

A decision tree is a decision support tool that uses a tree-like model of decision-making process and the possible consequences. It covers event

outcomes, resource costs, and utility of decisions. Decision Trees resemble an algorithm or a flowchart that contains only conditional control statements.

A decision tree is drawn upside down with the root node at top. Each decision tree has 3 key parts: a root node, leaf nodes, branches.

In a decision tree[11], each internal node represents a test or an event. Say, a heads or a tail in a coin flip. Each branch represents the outcome of the test and each leaf node represents a class label — a decision taken after computing all attributes. The paths from root to leaf nodes represent the classification rules.

Decision trees can be a powerful machine learning algorithm for classification and regression. Classification tree works on the target to classify if it was a heads or a tail. Regression trees are represented in a similar manner, but they predict continuous values like house prices in a neighborhood.

The best part about decision trees:

1) Handle both numerical and categoric data

2) Handle multi-output problems

3) Decision trees require relatively less effort in data preparation

4) Nonlinear relationships between parameters do not affect tree performance

Real life applications of Decision Trees:

1) Selecting a flight to travel

2) Predicting high occupancy dates for hotels

3) Number of drug stores nearby was particularly effective for a client X

4) Cancer vs non-cancerous cell classification where cancerous cells are rare say 1%

5) Suggest a customer what car to buy

1.1.2.1.4 RANDOM FOREST

Random Forests in machine learning is an ensemble learning technique about classification, regression and other operations that depend on a multitude of decision trees at the training time. They are fast, flexible, represent a robust approach to mining high-dimensional data and are an extension of classification and regression decision trees we talked about above.

A random forest should have a number of trees between 64–128 trees.

Ensemble learning, in general, can be defined as a model that makes predictions by combining individual models. The ensemble model tends to be more flexible with less bias and less variance.

Ensemble Learning has two popular methods as:

1) Bagging[11]: Each individual tree to randomly sample from the dataset and trained by s random subset of data, resulting in different trees

2) Boosting: Each individual tree /model learns from mistakes made by the previous model and improves

Random forest run times are quite fast. They are pretty efficient in dealing with missing and incorrect data. On the negatives, they cannot predict beyond the defined range in the training data, and that they may over-fit data sets that are particularly noisy.

Real life applications of Random Forests

1) Fraud detection for bank accounts, credit card

2) Detect and predict the drug sensitivity of a medicine

3) Identify a patient's disease by analyzing their medical records

4) Predict estimated loss or profit while purchasing a particular stock.

1.1.2.1.5 KNN

K- nearest neighbor (KNN) is a simple supervised machine learning algorithm that can be used to solve both classification and regression problems.

KNN stores available inputs and classifies new inputs based on a similar measure i.e. the distance function. KNN has found its major application in statistical estimation and pattern recognition.

KNN works by finding the distances between a query and all inputs in the data. Next, it selects a specified number of inputs, say K, closest to the query. And then it votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

The KNN[11] Algorithm:

1) Load the data

2) Initialize *k* to a chosen number of neighbors in the data

3) For each example in the data, calculate the distance between the query example and the current input from the data

4) Add that distance to the index of input to make an ordered collection

5) Sort the ordered collection of distances and indices in ascending order grouped by distances

6) Pick the first K entries from the sorted collection

7) Get the labels of the selected K entries

8) If regression, return the mean of the K labels; If classification, return the mode of the K labels

Real world applications of KNN:

1) Fingerprint detection

2) Forecasting stock market

3) Currency exchange rate

4) Bank bankruptcies

5) Credit rating

6) Loan management

7) Money laundering analyses

8) Estimate the amount of glucose in the blood of a diabetic person from the IR absorption spectrum of that person's blood.

9) Identify the risk factors for a cancer based on clinical & demographic variables.

1.1.2.1.6 SVM

SVM stands for Support Vector Machines. Machine learning largely involves predicting and classifying data. To do so, have a set of machine learning algorithms to implement depending on the dataset. One of these ML algorithms is SVM. The idea being simple: create a line or a hyperplane which separates the data into multiple classes.

Support Vector Machine[11] (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. SVM transforms your data base on it, finds an optimal boundary between the possible outputs.

Support Vector Machine performs classification by finding the hyperplane that maximizes the margin between the two classes.The vectors that define the hyperplane are called the support vectors.

The SVM Algorithm:

1) Define an optimal hyperplane with a maximized margin

2) Map data to a high dimensional space where it is easier to classify with linear decision surfaces

3) Reformulate problem so that data is mapped implicitly into this space

Real Life Applications of SVM:

1) Face detection — classify between face and non-face areas on images

2) Text and hypertext categorization

3) Classification of images

4) Bioinformatics — protein, genes, biological or cancer classification.

5) Handwriting recognition

6) Drug Discovery for Therapy. (In recent times, SVM has played a very important role in cancer detection and its therapy with its application in classification).

1.1.2.1.7 NAÏVE BAYES

Naive Bayes is super effective, commonly-used machine learning classifier. Naive Bayes is in its own a family of algorithms including algortihms for both supervised and unsupervised learning.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

$$P(A/B) = [P(B/A)*P(A)] / P(B)$$

Naive Bayes (NB) is naive because it makes the assumption that attributes of a measurement are independent of each other. We can simply take one attribute as independent quantity and determine proportion of previous measurements that belong to that class having the same value for this attribute only.

Naive Bayes is used primarily to predict the probability of different classes based on multiple attributes. It is mostly used in text classification while mining the data. If you look at the applications of Naive Bayes[11],

the projects you always wanted to do can be best done by this family of algorithms.

Real world applications of Naive Bayes

1) Classify a news article about technology, politics, or sports
2) Sentiment analysis on social media
3) Facial recognition softwares
4) Recommendation Systems as in Netflix, Amazon
5) Spam filtering

1.1.2.2 UNSUPERVISED LEARNING

Unsupervised learning is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown. The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the main attributes that separate customer segments from each other. Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition. These algorithms are also used to segment text topics, recommend items and identify data outliers.

Unsupervised learning is that algorithm where you only have to insert/put the input data (X) and no corresponding output variables are to be put.

The major goal for the unsupervised learning is to help model the underlying structure or maybe in the distribution of the data in order to help the learners learn more about the data.

These are termed as unsupervised learning because unlike supervised learning which is shown above there are no correct answers

and there is no teacher to this. Algorithms are left to their own devices to help discover and present the interesting structure that is present in the data.

Unsupervised learning problems can even be grouped ahead into clustering and association problems.

1) Clustering: A clustering is that problem which indicates what you want to discover and this helps in the inherent groupings of the data, such as grouping the customers based on their purchasing behavior.

2) Association: An association rule is termed to be the learning problem. This is where you would be discovering the exact rules that will describe the large portions of your data. Example: People who buy X are also the one who tends to buy Y.

Some common algorithms are:

1) K-means for clustering problems
2) Apriori algorithm for association rule learning problems
3) Principal Component Analysis
4) Singular Value Decomposition
5) Independent Component Analysis

1.1.2.2.1 K-means CLUSTERING

K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The *'means'* in the K-means refers to averaging of the data; that is, finding the centroid.

K-means algorithm starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either the centroids have stabilized or a defined number of iterations have been achieved.

The K-means clustering algorithm*:*

1) Specify the number of clusters *K*.

2) Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement

3) Keep iterating until the centroids are stabilized

4) Compute the sum of the squared distance between data points and all centroids

5) Assign each data point to the closest cluster (centroid)

6) Compute the centroids for the clusters by taking the average of the data points that belong to each cluster.

Real World applications of K-means Clustering

1) Identifying fake news

2) Spam detection and filtering

3) Classify books or movies by genre

4) Popular transport routes while town planning

1.1.2.2.2 APRIORI ALGORITHM FOR ASSOCIATION RULE LEARING PROBLEMS

Apriori[12] is considered an algorithm for frequent itemset mining and association rule learning over transactional databases. It proceeds just by identifying the frequent individual items in the database and then extending them to larger and larger item sets. The observation is, for as long as those itemsets appear sufficiently often in the database. The frequent itemsets that were determined by Apriori can be later used to determine about the association rules which highlights all the general trends that are being used in the database: this has got applications that fall in the domains such as the market basket analysis.

1.1.2.2.3 PRINCIPAL COMPONENT ANALYSIS

The main idea which falls behind the principal component analysis (PCA) is to help in reducing the dimensionality of the dataset which consists of many variables, that are always correlated with each other, either in a heavy or light manner, while retaining the variation which is present in the dataset, up to its maximum extent. The same thing is repeated and done by transforming and bringing the variables to a whole new set of variables, which are called the principal components (or simply, the PCs) and are even termed to be orthogonal, ordered in such a way that the retention of variation which is present in the original variables can be decreased as we try to move down in the proper order. So, by following this particular way, the 1st principal component retains the most and maximum variation that was earlier present in the original components. The principal components are basically known to be the eigenvectors of a covariance matrix, and hence they are even called the orthogonal.

Most importantly, the dataset which is based on what the PCA techniques are to be used and must be scaled. The result also turns out to be sensitive based on the relative scaling. As a layman, it can be termed as a method of summarizing data. Just imagine having some wine bottles on your dining table. Each wine would be described only by its attributes, that are like colour, age, strength, etc. But eventually, redundancy will arise maybe because many of them would be measured based on the related properties.

Principal component analysis might not be the best candidate in the algorithm category, but it definitely is super-useful as a machine learning technique.

Principal Component Analysis (PCA) is an unsupervised, statistical technique primarily used for dimensionality reduction by feature extraction in machine learning.

When we talk about high-dimensionality, it means that the dataset has a large number of features. and that requires a large amount of memory and computational power.

PCA uses orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables.It is used to explain the variance-co variance structure of a set of variables through linear combinations. It is a also the most widely used tool in exploratory data analysis and predictive modeling.

The idea behind PCA is simply to find a low-dimension set of axes that summarize data. Say for example, we have a dataset composed by a set of car properties; size, color, number of seats, number of doors, size of trunk, circularity, compactness, radius… However, many of these features will indicate the same result and therefore can be redundant. We as smart technologists should try to remove these redundancies and describe each car with fewer properties, making the computation simple. This is exactly what PCA aims to do.

PCA[12] does not take information of attributes into account. It concerns itself with the variance of each attribute because the presence of high variance would indicate a good split between classes, and that's how we reduce the dimensionality. PCA never just considers some while discards others. It takes the attributes into account statistically.

Real world applications of PCA:

1) Optimize power allocation in multiple communication channels
2) Image Processing
3) Movie recommendation system

1.1.2.2.4 Singular Value Decomposition

In linear algebra, you can call the singular-value decomposition (SVD) as a factorization of maybe real or complex matrix. It is the generalization of the eigendecomposition, that is the origin of a positive semidefinite normal matrix is done somewhere over here. It has many useful applications that are signal processing and are into statistics.

The singular-value decomposition[12] can be computed easily by making the use of the following observations:

- The left-singular vectors of M are considered to be a set of orthonormal eigenvectors of MM^*.
- The right-singular vectors of M are actually the set of orthonormal eigenvectors of M^*M.
- The non-zero singular values of M (that are found on the diagonal entries of Σ) are considered to be the square roots of the non-zero eigenvalues of both M^*M and MM^*.

Applications that help to employ the SVD include computing of the pseudoinverse, the least squares fitting of data, multivariable control, matrix approximation, and determining the rank, range and null space of a matrix.

1.1.2.2.5 INDEPENDENT COMPONENT ANALYSIS:

Independent component analysis (ICA), it is considered to be a statistical and computational technique. It helps to bring our or in revealing hidden factors that underlie in the sets of random variables, measurements, or signals.

ICA helps to define a generative model. This model stands for the observed multivariate data. It is typically recognized in the form of a large database of samples. Well, In the model, the data variables are assumed to

be the linear mixtures of few less known or you can call it as unknown latent variables, and even the mixing system is also unknown. Then comes the latent variables. These variables are actually assumed to be the nongaussian. They are even the mutually independent ones. These could be termed as the independent components belonging in the category of the observed data. These independent components, also termed as the sources or factors, can be found by the ICA.

ICA, the term is basically superficially related to the principal component analysis and then to the factor analysis. ICA is considered and supposedly it is a much more powerful technique[12]. Still, however this would be always capable of finding the underlying factors. It can even be the sources if possible by any chance, if these classic methods fail completely anyhow.

The data which is analyzed by the ICA could be originating from various kinds of application fields, this could be including digital images, the document databases, the economic indicators and then the psychometric measurements. In many cases, these measurements are given to be considered as a set of parallel signals or time series; the term blind source separation is then used in this to characterize this problem. Typical examples are actually the mixtures of simultaneous speech signals that have been picked up by several microphones, these are the brain waves that is recorded by multiple sensors and then the interfering radio signals that arriving at a mobile phone, or maybe the parallel time series which is obtained from performing some industrial process.

1.1.2.3 SEMI SUPERVISED LEARNING

In the previous two types, either there are no labels for all the observation in the dataset or labels are present for all the observations. Semi-supervised learning falls in between these two. In many practical situations, the cost to label is quite high, since it requires skilled human

experts to do that. So, in the absence of labels in the majority of the observations but present in few, semi-supervised algorithms are the best candidates for the model building. These methods exploit the idea that even though the group memberships of the unlabeled data are unknown, this data carries important information about the group parameters.

Semisupervised learning[13] is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data (because unlabeled data is less expensive and takes less effort to acquire). This type of learning can be used with methods such as classification, regression and prediction. Semisupervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. Early examples of this include identifying a person's face on a web cam.

1.1.2.4 REINCEFORCEMENT LEARNING

Reinforcement learning is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

Reinforcement Learning[13] is a type of Machine Learning, and thereby also a branch of Artificial Intelligence. It allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance. Simple reward

feedback is required for the agent to learn its behavior; this is known as the reinforcement signal.

There are many different algorithms that tackle this issue. As a matter of fact, Reinforcement Learning is defined by a specific type of problem, and all its solutions are classed as Reinforcement Learning algorithms. In the problem, an agent is supposed decide the best action to select based on his current state. When this step is repeated, the problem is known as a Markov Decision Process.

In order to produce intelligent programs (also called agents), reinforcement learning goes through the following steps:

1) Input state is observed by the agent.

2) Decision making function is used to make agent perform action.

3) After the action is performed, the agent receives reward or reinforcement from the environment.

4) The state-action pair information about the reward is stored.

1.2 MOTIVATION FOR WORK

Our motivation was to get the accurate result of the weather at that point of time. There were many applications and Google weather reports for predicting the weather but we wanted to forecast weather using image processing. All the applications need internet connectivity and Global Positioning System to say the weather at that place and at that time. Those are used to get the information or position of the user from the satellite and gives the resultant weather from it. Every individual can get the weather just by taking a single digital photo of the cloud and sky at which they are standing. This idea motivated us to do this project.

1.3 PROBLEM STATEMENT

To predict the weather at a particular area at any particular time with the help of the image of sky which should not contain any other objects other the sky and clouds. Even in remote areas where people don't have access to internet should also be able to get weather forecast. To be able to solve the above problem, we are doing this project.

1.4 ORGANIZATION OF THESIS

In this document, chapter 2 consists about literature survey. The literature survey tells about the research done to work on the project. All the details about the papers, websites on which the research work is done in order to work on the project is provided in the literature survey. In chapter 3, we discuss about the various methodologies used in the project. In chapter 4, the details about experimental analysis is discussed. The experimental analysis includes sample code, result screenshots for a tested input image. In the next chapter we give the conclusion about the project and also provide information if the project can be implemented further or not. In the final chapter we provide all the references for this project.

2. LITERATURE SURVEY

2.1 Weather Forecasting using satellite image processing and ANN:

The interpretation of satellite weather imagery has generally required the experience of a well-trained meteorologist. However, it is not always possible, or feasible to have an expert meteorologist on hand when such interpretation is desired. Therefore, the availability of an automated interpretation system would be quite desirable. Also, to take advantage of this available data in a reasonable and useful time increment, the system must be efficient and have low implementation cost. There are 3 main types of satellite images available - Visible, Infrared and Water Vapor[2]. Visible images are obtained only during the day. They are used to determine the thickness of the clouds. Infrared images are obtained using special infrared sensors. The major advantage of this type is that it can be obtained even during night. It can be used to measure temperature of cloud top. Water Vapor images indicate the moisture content or humidity. The brighter areas tend to have high chances of rainfall.

The author uses satellite images for the detection of weather at that moment of time. This paper uses the strategy which uses ICA/Fast ICA Algorithm that is proposed by Wang Yongqi and Du Huadong i.e. Studies on Cloud Detection of Atmosphere Remote Sensing using ICA algorithm. In this algorithm three types of images are considered as input. For the separation of image normalization is done to the un-mixing matrix obtained from the input which is used to segment the clouds. Considering the image, in this paper 2 phases are done: 1)Image processing phase where image segmentation is done for the region of interest and using the normalization cloud cover over the required region is calculated in percentage. In this paper, AABT is designed to provide an accurate and fast segmentation. Those results are potentially enough accurate for cloud cover percentage calculation. 2) machine learning phase where dataset is mapped with the calculated percentage obtained in image processing phase and the weather prediction is done by artificial neural network. The unique combination

of NAR and NARX neural network is used which produces positive result and accurate prediction to a very good extent.

For an automated weather satellite image interpretation system, one of the key steps is image segmentation. In this process, significant cloud features are extracted from the image and prepared for the next step in the process. AABT is designed to provide a fast and accurate method of image segmentation which is simple to implement as well. The segmentation results are provided quickly and with potentially enough accuracy to be integrated into a complete automated weather interpretation system or for cloud cover estimation. Furthermore, in case of the neural network model it can be successfully concluded from the above results that this unique combination of NAR and NARX neural network produce a positive result and the prediction is accurate to a good extent although there is always there lies a vast possibility of an error currently known as weather phenomenon[2] are such that the features required to beincorporated to create an extremely efficient model are very high, varied and in many cases incalculable. Although efforts in this area will always develop the scientific community as well as the world.

2.2 Machine learning applied to weather forecasting:

Weather forecasting is the task of predicting the state of the atmosphere at a future time and a specified location. Traditionally, this has been done through physical simulations in which the atmosphere is modelled as a fluid. The present state of the atmosphere is sampled, and the future state is computed by numerically solving the equations of fluid dynamics and thermodynamics. However, the system of ordinary differential equations that govern this physical model is unstable under perturbations, and uncertainties in the initial measurements of the atmospheric conditions and an incomplete understanding of complex atmospheric processes restrict the extent of accurate weather forecasting to a 10 day period, beyond which weather forecasts are significantly unreliable.

Machine learning, on the contrary, is relatively robust to perturbations and doesn't require a complete understanding of the physical processes that govern the atmosphere. Therefore, machine learning may represent a viable alternative to physical models in weather forecasting.

Two machine learning algorithms were implemented: linear regression and a variation of functional regression. A corpus of historical weather data for Stanford, CA was obtained and used to train these algorithms. The input to these algorithms was the weather data of the past two days, which include the maximum temperature, minimum temperature[3], mean humidity, mean atmospheric pressure, and weather classification for each day. The output was then the maximum and minimum temperatures for each of the next seven days.

In this paper, details of weather for the past 2 days is considered. Those details are considered as input and performing linear regression and variation of functional regression, output is obtained. The output is weather for next 10 days. Generally the classification of weather gives 9 classes: clear, scattered clouds, partly cloudy, snow, thunder strom, rain, overcast, fog, mostly cloudy. The dataset considered, classified all those into 3 classes: moderate cloudy, very cloudy, precipitation. The least mean square error for the linear regression and variation on functional regression is calculated and learning curves are drawn in this paper. Linear regression is low biased with high variance model whereas functional is exactly opposite to it. Collection of more data can improve the linear regression model. Hence the author suggests to consider 4 to 5 days of data as input to the model.

Both linear regression and functional regression were outperformed by professional weather forecasting services, although the discrepancy in their performance decreased significantly for later days, indicating that over longer periods of time, our models may outperform professional ones. Linear regression proved to be a low bias, high variance model whereas functional regression proved to be a high bias, low variance model. Linear regression is inherently a high variance model as it is unstable to outliers, so one way to improve the linear regression model is by collection of more data.

Functional regression, how- 5 ever, was high bias, indicating that the choice of model was poor, and that its predictions cannot be improved by further collection of data. This bias could be due to the design choice to forecast weather based upon the weather of the past two days, which may be too short to capture trends in weather that functional regression requires. If the forecast were instead based upon the weather of the past four or five days, the bias of the functional regression model could likely be reduced. However, this would require much more computation time along with retraining of the weight vector w, so this will be deferred to future work.

2.3 Analysis on various techniques for weather forecasting:

1) Support vector machines

To predict the maximum temperature of a required location Support Vector Regression (SVR) is used. It performs better than MLP which is trained with back propagation algorithms as it minimizes the upper bound on generalization error. By selecting proper parameters it can replace neural networks based models for applications of weather prediction.

2) Time Series Analysis for Weather Forecasting

Data groups and data variables in the specified time are captured by Time Series Analysis. By comparing actual and predicted values of temperature, the forecasting reliability was evaluated. The results show that important tool for temperature forecasting is network.

3) Prediction of Weather by using Back Propagation Algorithm

Wind, humidity, rainfall and temperature are the parameters recorded using sensors. Using these sensors weather forecasting and processing information is transferred. It classifies, compares and predicts the change in other weather parameters by changing any one parameter

value that those sensors recorded. A 3 layered neural networks[4] trained with the existing dataset to develop a relation among the parameters of weather that are non linear.

4) Fuzzy Logic Based Rainfall Prediction model:

Two components are made in a developed fuzzy logic model where one is knowledge based and the other is fuzzy reasoning or decision making. Using fuzzification and defuzzification operations outputs are predicted compared with actual rainfall data. A fuzzy model that is well developed is capable of handling the data that is scattered and shows flexibility in modelling weak input and output variable relationship.

2.4 Weather forecasting using data mining research based on cloud computing

Weather Prediction is the application of science and technology to predict atmospheric conditions ahead of time for a particular region. Prediction is one of the basic goals of Data Mining. Data Mining is to dig out knowledge and rules, which are hidden and unknown. User may be interested in or has potential value for decision-making from the large amounts of data. Such potential knowledge and rules can reveal the laws between the data. There are many kinds of technical methods of data mining, which mainly include: association rule mining algorithm, decision tree classification algorithm, clustering algorithm and time series mining algorithm, etc. How to store, manage and use these massive meteorological data, discover and understand the law and knowledge of the data, to contribute to weather forecasting completely and effectively has attracted more and more Data Mining researcher's attention. This article constructs the Weather Forecasting platform, using data mining for meteorological forecast and the forecast results are analysed.

Before Cloud computing[3] has improved the efficiency of data storage, delivery, and dissemination across multiple platforms and applications, allowing

easier collaboration and data sharing, including data processing and distribution systems that disseminate key weather forecasting, severe weather warning, and climate information. Data mining techniques and forecasting applications are very much needed in the cloud computing paradigm. In this study, data mining in Cloud Computing allows weather forecasting and data storage, with assurance of efficient, reliable and secure services for their users. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that will reduce the costs of infrastructure and storage.

A modern method is developed which is service oriented architecture for the weather information systems that forecasts weather using data mining techniques. The method uses Artificial Neural Network and Decision tree Algorithms and meterological data collected in specific time. It presents the best results for generating classification rules for the mean weather variables. The model predicts temperature, rainfall and wind speed. Cloud computing reduces the cost of infrastructure and storage as it ensures secure reliable and efficient services for the user.

The implementation of data mining approach to solve the wind forecasting problems for wind farm production, in particular, for predicting wind speed. The data mining prediction algorithm-ARIMA time series prediction algorithm is also integrated into the system. The platform has the ability of mass storage of meteorological data, efficient query and analysis, weather forecasting and other functions. In this study we also adapted the method of Artificial Neural Networks, it can detect the relationships between the input variables and generate outputs based on the observed patterns inherent in the data without any need for programming or developing complex equations to model these relationships. An artificial neural network (multi-layer perceptron) was applied and several simulations have been conducted for comparison purposes. ANN's can detect the relationships between weather parameters and use these for future prediction. Weather conditions[5] are important to climatic change studies because the variation in weather conditions in term of temperature and wind speed can be

studied using these data mining techniques. ANNs are implemented, in order to compare their effectiveness in changing the network topology and the training mode. The results obtained from real data are based on time series of meteorological data provided by the Dalian Meteorological Bureau. The test cases pointed out that the proposed approach gives a very interesting performance of the implemented network and shows good performance in term of MSE. For future perspective, there is still significant potential for improvement in weather forecasting by using ANN model, through introducing climate change and global warming variables, in order to forecast more realistic weather parameters.

2.5 Cloud image analysis and classification

Clouds are classified by their shape, temperature, color, density, spectral clustering analysis, training of rule based systems or neural networks and image processing techniques. Classification of clouds are highly time dependent because temperature, type of cloud dependence changes in different latitudes through different seasons. Different types of cloud[5] are present out of which three basic cloud forms are the Cirrus, Stratus, and Cumulus.

These forms are further refined into 10 other types based on their height and texture.

1. Cirrus:

   Thin, white and feathery appearance and mostly white patches or narrow bands.

2. Cirrocumulus:

   Thin white bands or ripples, sheet, or layered of clouds without shading.

3.Cirrostratus:

   High, milky white like appearance. They are transparent, whitish veil clouds with a fibrous (hair-like) or smooth appearance.

4. Altocumulus:

Bumpy rounded masses, cotton ball appearance, white and/or gray patch sheet or layered clouds.

5. Altostratus:

Transparent blue/gray clouds sheets or fibrous clouds that totally or partially cover the sky.

6. Nimbostratus:

They are continuous rain cloud also known as storm cloud.

7. Stratocumulus:

Gray or whitish layer with sheet, or layered clouds which almost always are dark.

8. Stratus:

Cover large portion of sky, thin, sheet-like, gray and thick.

9. Cumulus:

Cauliflower like appearance with bulging upper parts.

10. Cumulonimbus:

The thunderstorm cloud, heavy and dense cloud in the form of a mountain or huge tower.

2.6 Existing system

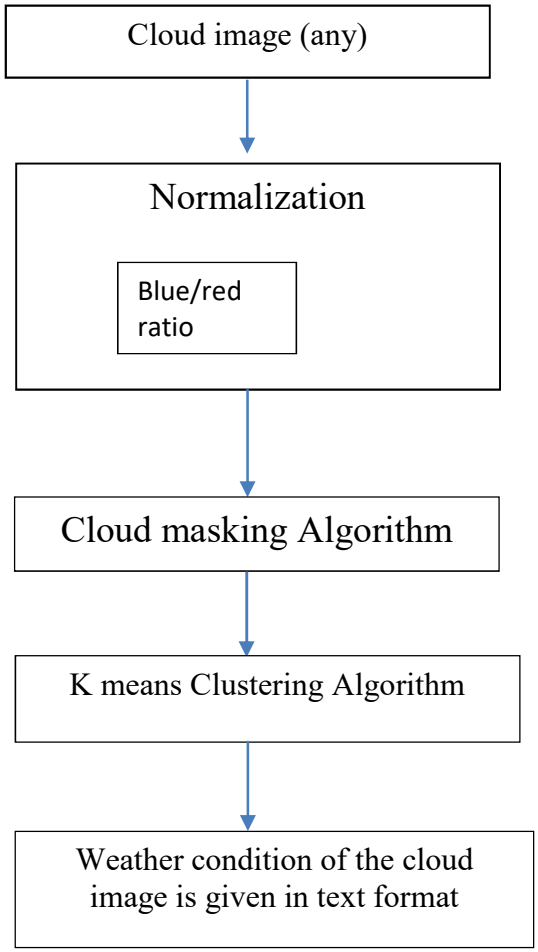
The existing system for weather forecasting mainly uses satellite images to predict the weather. Most of the websites which give weather forecast with the help of GPS as well as satellite images data to predict the weather. Very rarely we find applications or websites predicting weather using digital image processing. These applications or websites use the satellite images in which we get the data about the cloud cover over any particular region and by using this data they predict the weather. They track the movement of the clouds for a particular time period and use that to predict weather for the future. This is the existing system for weather forecasting.

3. METHODOLOGY

3.1 PROPOSED SYSTEM

Weather forecasting can also be done by using satellite images but acquiring the satellite images is more difficult and would even cost high. Even predicting using the satellite images needs more technology. So, we are using digital image processing techniques which processes the images of the sky like normalization, cloud masking algorithm and k-mean algorithm.

Architecture :



Every system should be divided into modules for better understanding and execution. Dividing into modules helps the programmer and client to work and use the system efficiently, respectively. If any system is not divided into modules and worked as a whole, then there comes a numerous errors. Even we find difficulty in correcting those errors. It is must and should to divide the total project into modules and work on each and every module independently to get effective results. Our total system is divided into three modules namely:

1. Normalization of Image
2. Cloud masking algorithm
3. K-means clustering

3.2 Normalization of Image

Pixel values for each and every pixel are considered. Pixel value consists of red, blue and green color's values. These values are extracted from the image with the help of pre-defined libraries in python. Now with the help of these pixels, we have to change the intensity range of the pixels to [0,1] and increase the intensity to get a clear distinction between the clouds and the sky. Hence the digital picture is normalized[1]. The input image can be of any digital image with the extension .jpg, .jpeg, .png. The output of this module would be a normalized image of the given digital image which seems likely to be a black and white or gray scale image.

We used different formulas to get the image normalized using the red, blue and green values of the each pixel. And the threshold value is generated by taking the mean of all the pixel values. The input image can be of any digital image with the extension .jpg, .jpeg, .png. And the size of the image must vary between 20kb to 20mb. The output of this module would be a normalized image of the given digital image which seems likely to be a black and white or gray scale image.

ALGORITHM:

Step 1:

Pixel values for each and every pixel are considered. Pixel value consists of red, blue and green color's values. These values are extracted from the image with the help of pre-defined libraries in python.

Step 2:

Now with the help of these pixels, we must change the intensity range of the pixels to [0,1] and increase the intensity to get a clear distinction between the clouds and the sky. Hence the digital picture is normalized.

Step 3:

The input image can be of any digital image with the extension .jpg, .jpeg, .png.

Step 4:

The output of this module would be a normalized image of the given digital image which seems likely to be a black and white or gray scale image.

Fig. 1 Original image of cloud

Fig. 2 Image after performing normalization

3.3 Cloud Masking Algorithm

Considering the threshold value generated in the normalization module, we differentiate the cloud from the sky[10] of the input image so that we can get the area of the cloud upon which we will perform further operations like feature extraction.

The data set is considered and normalization[9] is done for each and every image in it. Then after getting the clouds separated from the image, mean point is derived from each cloud. Based on those mean points, clouds are separated. Comparing the input image threshold value and mean value of the cloud, it would be pushed into that category of cloud for which its values coincides.

The output of this module would be like in Fig 3.

Step 1:

After normalization, a mean value is generated by adding all the pixel values and by dividing it by the total no of pixels. With the help of this mean value we differentiate the clouds from the input image.

Step 2:

Now extract the feature of the cloud part by again finding the mean value of the cloud area which will be used a feature in the next process.

Step 3:

This process is done for all the images in the dataset so that we get features of all images which will be used to cluster the images into groups.

Figure 3.Image obtained after performing cloud masking algorithm

3.4 K means clustering

We considered clustering because for classification there would be less no of classes. But we considered ten types and hence we considered clustering rather than classification. Here we considered the clusters of the clouds as we would divide the image based on the cloud mean point.

In this the clouds are divided into 10 clusters[6]. The clusters are:

1. Cirrostratus
2. Cirrus
3. Cirrocumulus
4. Altocumulus
5. Altostratus
6. Stratus
7. Stratocumulus
8. Nimbostratus
9. Cumulonimbus
10. Cumulus

Among these the cirrostratus, altostratus, cirrocumulus, cirrus denotes sunny day. Nimbostratus and Cumulonimbus denotes rainy day. And cumulus, stratus, altocumulus and stratocumulus denotes cloudy day.

All the classification[6] can be done depending on the mean threshold value. From the dataset after applying normalization and cloud masking algorithm we can get a threshold value for each and every cloud cluster. Based on that value i.e. that threshold value the input image is classified into one of the ten clusters. Then based on that cluster type the type of cloud can be predicted. After that based on the type of the cloud weather can be forecasted as per the above argument.

4. EXPERIMENTAL ANALYSIS AND RESULTS

4.1 SYSTEM CONFIGURATION:

4.1.1 Software requirements:

1) Python :

Python is an interpreted high level and general purpose programing language created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

2) PyCharm :

PyCharm is the only integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as Data Science with Anaconda. PyCharm is cross-platform, with Windows, macOS and Linux versions. The Community Edition is released under the Apache License and there is also Professional Edition with extra features – released under a proprietary license.

3) OpenCV :

OpenCV (Open source computer vision) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-

source BSD license. OpenCV supports some models from deep learning frameworks like TensorFlow, Torch, PyTorch (after converting to an ONNX model) and Caffe according to a defined list of supported layers. It promotes OpenVisionCapsules. which is a portable format, compatible with all other formats.

4) Numpy :

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

4.1.2 Hardware requirements

1) Processor: 64 bit, quad-core, 2.5 GHz minimum per core

2) Ram: 4 GB or more

3) Hard disk: 20 GB of available space or more.

4) Display: Dual XGA (1024 x 768) or higher resolution monitors

4.2 SAMPLE CODE

```
import cv2
import numpy as np
import os
```

```
import pandas as pd
import random as rd,math
import sys

path='C:\\Users\\LENOVO\\Desktop\\images dataset\\'
files=[]
meansl=[]
m=[]
M=[]
c=0
for r,d,f in os.walk(path):
    for file in f:
        if('.jpeg' in file or '.jpg' in file):
            files.append(os.path.join(r,file))
            c+=1
#print(files)
for input_file in files:
    f=cv2.imread(input_file)
    h, w, bpp = np.shape(f)
    for py in range(0, h):
        for px in range(0, w):
            x = float(f[py][px][0])
            y = float(f[py][px][1])
            z = float(f[py][px][2])
            f[py][px] = x * 0.2126 + y * 0.0722 + z * 0.7152
    f = (255 / 1) * (f / (255 / 1)) ** 2
    h, w, bpp = np.shape(f)
    n = np.zeros([h, w, bpp], dtype=np.uint8)
    c = 0
    sum = 0
    for py in range(0, h):
```

```
            for px in range(0, w):
                sum += f[py][px][0]
                c += 1
        for py in range(0, h):
            for px in range(0, w):
                if f[py][px][0] > (sum / c):
                    n[py][px] = f[py][px]
                    f[py][px] = 255

        sum1 = 0
        z = 0
        for py in range(0, h):
            for px in range(0, w):
                if n[py][px][0] != 0:
                    sum1 += n[py][px][0]
                    z += 1
        mean = sum1 / z
        dict={}
        dict1={}
        dict[input_file]=mean
        dict1=mean
        print(dict1)
        meansl.append(dict1)
    print(meansl)
    M.append(max(meansl))
    m.append(min(meansl))

    def InitializeMeans(meansl, k, m, M):
        f=1;#no. of features
        means=[[0 for i in range(f)] for j in range(k)];
```

```
    for item in means:
        for j in range(len(item)):
            item[j]= rd.uniform(m[j]+1,M[j]-1);
    return means;

def EuclideanDistance(x,y):
    S=0;
    for i in range(1):
        S += math.pow(x-y, 2);
    return math.sqrt(S);



def UpdateMean(n, mean, item):
    for i in range(len(mean)):
        m2 = mean[i];
        m2 = (m2 * (n - 1) + item) / float(n);
        mean[i] = round(m2, 3);

    return mean;



def CalculateMeans(k, items, maxIterations=100000):
    cMin=m;
    cMax=M;

    # Initialize means at random points
    means = InitializeMeans(meansl, k, cMin, cMax);

    # Initialize clusters, the array to hold
    # the number of items in a class
    clusterSizes = [0 for i in range(len(means))];
```

```
# An array to hold the cluster an item is in
belongsTo = [0 for i in range(len(meansl))];

# Calculate means
for e in range(maxIterations):

    # If no change of cluster occurs, halt
    noChange = True;
    for i in range(len(meansl)):

        item = meansl[i];

        # Classify item into a cluster and update the
        # corresponding means.
        index = Classify(means, item);

        clusterSizes[index] += 1;
        cSize = clusterSizes[index];
        means[index] = UpdateMean(cSize, means[index], item);

        # Item changed cluster
        if (index != belongsTo[i]):
            noChange = False;

        belongsTo[i] = index;

        # Nothing changed, return
    if (noChange):
        break;
```

```
    return means;


def Classify(means, item):
    minimum = sys.maxsize;
    index = -1;

    for i in range(len(means)):

        dis = EuclideanDistance(item, means[i]);

        if (dis < minimum):
            minimum = dis;
            index = i;

    return index;


def FindClusters(means, meansl):
    clusters = [[] for i in range(len(means))];  # Initialize clusters

    for item in meansl:
        index = Classify(means, item);

        clusters[index].append(item);

    return clusters;


means=CalculateMeans(4,meansl)
means.sort()
```

```
print(means)
g=FindClusters(means,meansl)
print(g)


m = cv2.imread("i32.jpg")

h,w,bpp = np.shape(m)
red=[]
blue=[]
green=[]

for py in range(0,h):
    for px in range(0,w):
        red.append(m[py][px][0])
        blue.append(m[py][px][1])
        green.append(m[py][px][2])

red_max=max(red)
blue_max=max(blue)
green_max=max(green)
red_min=min(red)
blue_min=min(blue)
green_min=min(green)

for py in range(0,h):
    for px in range(0,w):
        x=float(m[py][px][0])
        y=float(m[py][px][1])
        z=float(m[py][px][2])
        '''r=(x-red_min)*(1/(red_max-red_min))
```

```
    b=(y-blue_min)*(1/(blue_max-blue_min))
    g=(z-green_min)*(1/(green_max-green_min))'''


    m[py][px]=x*0.2126+y*0.0722+z*0.7152


m = (255 / 1) * (m / (255 / 1)) ** 2
h,w,bpp = np.shape(m)
n=np.zeros([h,w,bpp], dtype=np.uint8)
c=0
sum=0
for py in range(0,h):
   for px in range(0,w):
      sum+=m[py][px][0]
      c+=1

for py in range(0,h):
   for px in range(0,w):
      if m[py][px][0]>(sum/c):
         n[py][px]=m[py][px]
         m[py][px]=255

sum1=0
z=0
for py in range(0,h):
   for px in range(0,w):
      if n[py][px][0]!=0:
         sum1+=n[py][px][0]
         z+=1
mean=sum1/z
```

```
print(mean)

for i in range(len(means)):
    if mean>means[i]:
        v=i;
if v!=len(means)-1:
    o1=means[v+1]-mean;
    o2=mean-means[v];
    if o1<o2:
        v=v+1;
    print(v)
else:
    v=len(means)-1;

print('current weather condition is:')
if v==0:
    print('CLEAR SKY')
elif v==1:
    print('SUNNY')
elif v==2:
    print('CLOUDY AND SUNNY')
elif v==3:
    print('CLOUDY WITH CHANCES OF RAIN')
```

4.3 SCREENSHOTS

4.3.1 Normalization using blue-red ratio:

Implementation of the total module is completed. The input image can be of any digital image with the extension .jpg, .jpeg, .png. And the size of the image must vary between 20kb to 20mb. The output of this

module would be a normalized image[9] of the given digital image which seems likely to be a black and white or gray scale image.

Given input:

Fig. 4 Original image

Observed output:

Fig.5 Image obtained after performing normalization

4.3.2. Cloud masking algorithm:

When coming to the status of implementation, code has been developed up to a certain extent where the cloud detection is performed. The feature extraction of the cloud[10] still needs to be done and also the training of dataset is yet to be started.

Cloud detection:

Fig6. Input image for cloud masking algorithm

Fig. 7 Image obtained after cloud masking algorithm.

4.3.3. K-means clustering:

K-means clustering is done on the dataset to cluster the images in the dataset into 4 clusters. This process is done by obtaining the means of the images which is the main feature considered for clustering. From the list containing means we consider 'k' random means as the initial mean[6] points. From these points we calculate distance between the other means with the help of Euclidean distance formula:

Sqrt (S += math.pow(x-y, 2))

After finding the distances, we update the randomly selected means based on the distances. This is an iteration process which is repeated until there is no change in the randomly selected means. Once the

process is complete we get the means of the clusters which is considered as the centroid of cluster. Based on these means we can find to which cluster the given input image belongs.

4.4 EXPERIMENTAL ANALYSIS/ TESTING:

4.4.1. Dataset:

The dataset we considered is named “HYTA”. It consists of various images of all types of clouds. We considered 4 clusters for all the types of clouds namely: clear sky, sunny, cloudy and sunny, rainy. For each type of cluster this HYTA dataset[14] consists of nearly 8 to 10 images. Every image consists of only plain sky with respective clouds and no other objects like buildings, trees and poles. In some images the sun might appear along with the sky and clouds. Along with this standard dataset we considered 4 different datasets. Every data set consists of more than 10 photos for each type of cloud. These datasets are considered to compare the outputs obtained and check the accuracy of the model developed. This comparison will be useful for the future development of the model.

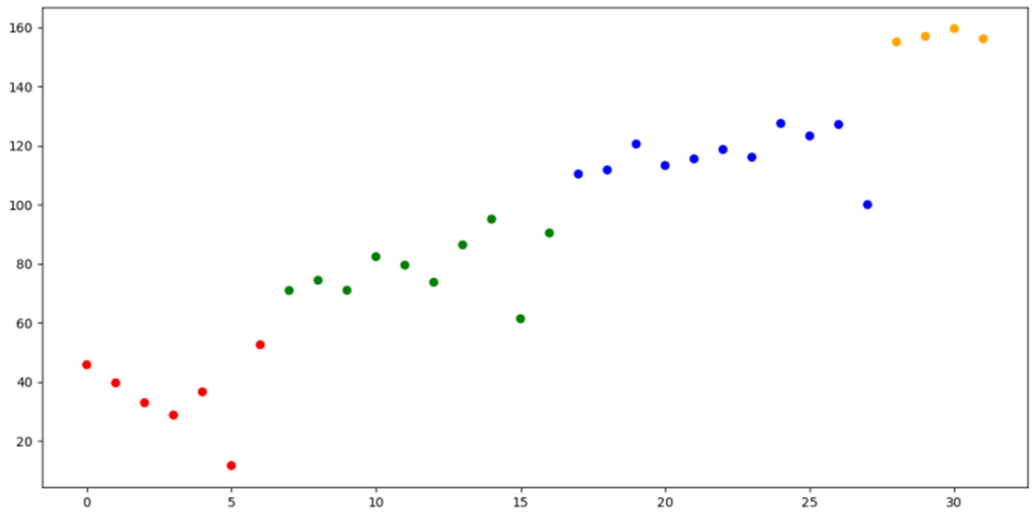


Fig. 8 Scatterplot graph[8] of different clusters formed after training the dataset

4.4.2. ANALYSIS

The final output obtained for the input image is given in a text format as shown in the figure below.

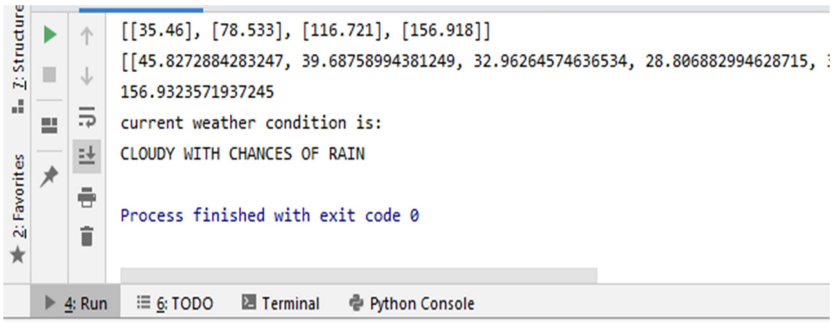


Fig.9 Weather forecast of the input image in text format

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

Generally, all the other weather forecasting applications and sources would give the weather report of that particular area. Using the GPS, location would be tracked and using satellite information, weather condition would be given at that place. The types of clouds the author considered, can deliberately give accurate condition of the weather. For now, the model can give the weather condition at that point of time.

5.2 Future Work

To get weather forecast for the next few days, we can modify our system by using different algorithms and use that as an extension to our current project.

REFERENCES

1. https://www.mathworks.com/matlabcentral/answers/422493-how-to-do-normalized- blue-red-ratio-operation-of-an-image
2. Kapadia Nilay S, Urmil Parikh in Nov 2016, Weather Forecasting using Satellite Image Processing and Artificial NueralNetworks. IJCSIS vol 14, No.11,
3. MinakshiGogoi and Gitanjali Devi (Oct-Dec 2015), Cloud image analysis for rainfall prediction, Advanced Research in EEE.
4. Cazorla, A., J. Olmo, and L. AladosArboledas, 2008: Development of a sky imager for cloud cover assessment. *J. Opt. Soc. Amer. A*, **25**, 29 39
5. Calbo, J., and J. Sabburg, 2008: Feature extraction from whole-sky ground-based images for cloud-type recognition. *J. Atmos. Oceanic Technol.*, **25**, 3–14
6. https://www.geeksforgeeks.org/k-means-clustering-introduction/
7. https://journals.ametsoc.org/doi/full/10.1175/JTECH-D-11-00009.1
8. https://www.science-emergence.com/Articles/How-to-create-a-scatter-plot-with-several-colors-in-matplotlib-/
9. https://in.mathworks.com/matlabcentral/answers/422493-how-to-do-normalized-blue-red-ratio-operation-of-an-image
10. https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13
11. https://towardsdatascience.com/top-10-algorithms-for-machine-learning-beginners-149374935f3c
12. https://www.newtechdojo.com/list-machine-learning-algorithms/
13. https://www.sas.com/en_in/insights/analytics/machine-learning.html
14. https://github.com/Soumyabrata/HYTA

Weather Forecasting using Satellite Image Processing and Artificial Neural Networks

Nilay S. Kapadia
Department of Computer Engineering
Sardar Vallabhbhai National Institute of Technology Surat
Surat 395007, Gujarat, India
nilayskapadia95@gmail.com

Urmil Parikh
Department of Computer Engineering
Sardar Vallabhbhai National Institute of Technology Surat
Surat 395007, Gujarat, India
urmil.parikh99@gmail.com

Dipti P. Rana
Department of Computer Engineering
Sardar Vallabhbhai National Institute of Technology
Surat 395007, Gujarat, India
dpr@coed.svnit.ac.in

The advent of new satellite imaging technologies has made satellite images more accessible. These images can be utilized for weather predictions. This work proposes a simple approach for weather prediction that relies on satellite images and weather data as inputs. The method is divided into two parts. The first part involves the use of image processing techniques such as image segmentation on the satellite images to extract the cloud cover. On basis of the cloud cover obtained, percentage cloud cover is calculated and this calculated percentage value is stored, which is later used in the second stage of the approach. The second part involves the use of the cloud cover percentage along with other inputs such as temperature, humidity and wind speed to train an artificial neural network. The weather prediction is done by artificial neural networks. Most of the current cloud extraction algorithms are quite complicated to implement and execution time is potentially slow. In this paper, we present a novel approach which is simple to implement, fast in execution and provides good results in tests.

Keywords-Satellite Images, Image Processing, Artificial Neural Networks

I. INTRODUCTION

The interpretation of satellite weather imagery has generally required the experience of a well-trained meteorologist. However, it is not always possible, or feasible to have an expert meteorologist on hand when such interpretation is desired. Therefore, the availability of an automated interpretation system would be quite desirable. Also, to take advantage of this available data in a reasonable and useful time increment, the system must be efficient and have low implementation cost. There are 3 main types of satellite images available - Visible, Infrared and Water Vapor. Visible images are obtained only during the day. They are used to determine the thickness of the clouds. Infrared images are obtained using special infrared sensors. The major advantage of this type is that it can be obtained even during night. It can be used to measure temperature of cloud top. Water Vapor images indicate the moisture content or humidity. The brighter areas tend to have high chances of rainfall.

In recent years the exponential increase in processing power has revived machine learning algorithms like artificial neural networks, linear and logistic regression. This has resulted in wide range development of machine learning algorithms for nearly every application, from handwriting recognition to solve crimes or predicting the stock market. Weather prediction using machine learning techniques is a field where much research has not been done. Predicting the weather is one of the most important and challenging aspects of remote sensing due to a large number of factors affecting it.

There has been significant progress in the area of remote sensing of satellite images using image processing methods. One of the strategies used for image retrieval and feature extraction is using fuzzy SOM strategy for satellite image retrieval and information mining projected by yo-ping hung, tsun-wei and li-jen kao[1]. They proposed a model for efficient satellite image retrieval and knowledge discovery. It has two major parts. First, it uses a computation algorithm for off-line satellite image feature extraction, image data representation and image retrieval. A self-organization feature is used to create a two-layer satellite image concept hierarchy. The events are stored in one layer and the corresponding feature vectors are categorized in the other layer. Another strategy proposed by Craig M. Wittenbrink et. al is Feature extraction of clouds from GOES satellite data for integrated model measurement visualization [2]. The paper suggests a de-correlating transformation to the spectral images using Karhunen-Loeve Transformation (KLT) (more properly known as the Hoteling transform). The KLT has been widely used in remote sensing for multispectral imagery, and is also known as principal component analysis. The principal components are obtained, and the first n are selected. The choice of n is a trade-off between low analysis complexity and accurate representation. The three main components are then mapped into a 3-D histogram. One of the more recent strategies is the use of ICA/Fast ICA Algorithm proposed by Du Huadong and Wang Yongqi is Studies on Cloud Detection of Atmospheric Remote Sensing Image using ICA Algorithm. In this strategy the 3 types of images (AVHRR) are used as input in the algorithm [3]. The un-mixing matrix obtained from it can be used to segment clouds from the image. To show the different object in the separated

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

component more clearly, the normalization is done to the separated image. A paper by Chiang Wei et al suggests a multi-spectral spatial convolution approach for real-time rainfall forecasting using geostationary weather satellite images [4]. The approach incorporates cloud-top temperatures of three infrared channels in a spatial convolution context. The kernel function of the multispectral spatial convolution equation is solved by the least squares method.

Initial studies in this area were done in 1998 in a paper titled "Localized Precipitation Forecasts from a Numerical Weather Prediction Model Using Artificial Neural Networks" by Robert J. Kuligowski and Ana P. Barros [5]. They proposed use of basic combination of back propagation neural network using a sigmoid function to predict data. Although most approaches in this area are restricted to the use of feed forward neural network. There have also been a few applications of different types of neural networks such as that by M.W Gardner and S.R Dorling in their paper titled "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences" [6]. Here the authors present a general introduction and discussion of recent applications of the multilayer perceptron in the atmospheric sciences especially weather prediction.

II. PROPOSED APPROACH

The proposed approach of solving the problem consists of two main phases, the initial phase is the image processing phase where image segmentation is used to segment the satellite image of the region of interest (the region whose weather has to be predicted). After segmenting the image cloud cover over the required region is extracted in the form of a percentage value.

The next phase is the machine learning phase where weather data (humidity, temperature, etc.) is combined with the cloud cover obtained from the satellite image. This combined table of different parameters is then fed into an artificial neural network for training the model.

III. SATELLITE IMAGE PROCESSING

The section involves a list of steps. The first step is to perform image segmentation, in which the cloud features of interest are extracted from the original satellite image. The following steps then involve feature analysis and interpretation. Here we wish to extract cloud cover from the images. The steps for cloud cover extraction are Image Segmentation, Region Separation using image cropping and Percentage cloud cover extraction.

A. Image Segmentation

For this part a simple algorithm is required which must achieve two goals. The first of which is to provide satisfactory image segmentation. The second goal is to provide an algorithm which is simple to implement and has relatively fast execution.

Adaptive Average brightness thresh-holding (AABT) looks promising in addressing these two goals. This method is based on four observations [7]:

1. Amplitude thresh-holding is simple to implement and provides quick processing.

2. When correct threshold level(s) are chosen amplitude thresh-holding is highly effective.

3. Clouds are usually the brightest objects in a weather satellite image.

When processing a cloud satellite image, AABT follows a series of steps. First, the image is divided into approximately equal sized quadrants. Second, for each quadrant an average brightness level is calculated. Next, using an average cut-off function a suitable cut-off threshold is determined for each of the quadrants and applied to each region separately. Finally, the complete image is produced by recombining the sub-regions [7].

For the third step of this algorithm, the average cut-off function is given by:

$$Cutoff = Avg.Brightness + f * (\ln(GMAX) - \ln(Avg.Brightness)) \quad (1)$$

Where:

ln() denotes the natural logarithm

G-MAX is the number of greyscale values in this case, G-MAX = 256

f is a multiplicative coefficient, determined empirically, in this case, f = 22.5

B. Region Seperation

This technique involves cropping out the region for which the prediction is to be done. This way a better view of the region is obtained. By observing the surrounding regions in the image weather conditions can be noted. Also processing of smaller images is faster and easier than large images. Although problem faced is that the resolution of image decreases as compared to the original image.

C. Cloud Cover Extraction and Percentage Calculation

The main aim of image processing is to obtain the cloud cover. The amount of cloud present in the region is determined by the cloud cover percentage. It serves as an important parameter for prediction of weather.

It is calculated by using this formula:

$$CloudCover(\%) = \frac{No.of\ nonzero\ pixels}{Total\ no.of\ pixels} * 100 \quad (2)$$

Here pixel represents the values of the image matrix.

From this, we get the values of cloud cover. These values are used as input parameters for training samples through an artificial neural network.

IV. PREDICTION USING ANN

Artificial neural networks (ANN) form the basis for a number of computational models based on the mammalian brain [8]. Instead of depending on linear correlative relationships

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

among a particular dataset, ANN is a form of machine learning in which the system learns to predict an output variable based on an input series. Data is processed by feeding it to a number of interconnected neurons which form synaptic connections. These connections follow a path from the input nodes through a hidden layer before ending on the output neurons. Each input and hidden neuron consist of statistical weights which are capable of adaptation, the exact parameters which are modified by an algorithm over the course of network training procedures. The weights form the synaptic connections between neurons which are activated during network creation. This form of computing has the ability to operate in a parallel format, similar to the human nervous system. Because neural networks do not depend on linear dependencies for learning, ANNs are capable of nonlinear modeling and therefore, provide an alternative approach to a number of theoretical and real-world problems.

The following sub-sections explain the two types of neural networks used i.e. Non-Linear Autoregressive (NAR) and Non-Linear Autoregressive Exogenous model (NARX).

A. Non-Linear Autoregressive Neural Network Model

This particular form of ANN is used in time series modeling in order to predict an outcome variable [y(t)] based on d past values of the outcome variable. The fit of predicted output values can be compared to the original target values using simple correlation coefficients, while an error term is also employed in order to further gauge predictive accuracy, usually presented as some form of mean or summed squared error term (target-output).

$$y(t)=f\ (y\ (t-1),\dots\ ,y(t-d)) \tag{3}$$

The other neural network model used is the Non-Autoregressive Neural Network model with exogenous input (NARX).

B. Non-Linear Autoregressive Exogenous Model

This particular form of ANN is used in time series modelling in order to predict an outcome variable [y(t)] based on 'd' past values of the outcome variable and current as well as 'd' past values of an external source of influence [x(t)] [9]. The accuracy of the predicted output values can be compared with the original target values using correlation coefficients, while an error term is also used in order to improve gauge predictive accuracy, usually presented as some forms of mean/summed squared error term (target-output).

$$y(t)=f(x(t-1),\dots,x(t-d),y(t-1),\dots,y(t-d)) \tag{4}$$

V. Implementation Process for Satellite Image Processing

To test the working and the performance of the method, it was applied on a dataset of images. This dataset was collected from the website of Indian Meteorological Department. The dataset consisted of Water vapor images. The images used were taken at a duration of every one hour of each day for the year 2015.

The methodology implemented on a sample satellite image has been shown below:

A. Image Segmentation

It is the first step of the method. The AABT algorithm is applied on the image. The result is as shown:

The methodology implemented on a sample satellite image has been shown below:

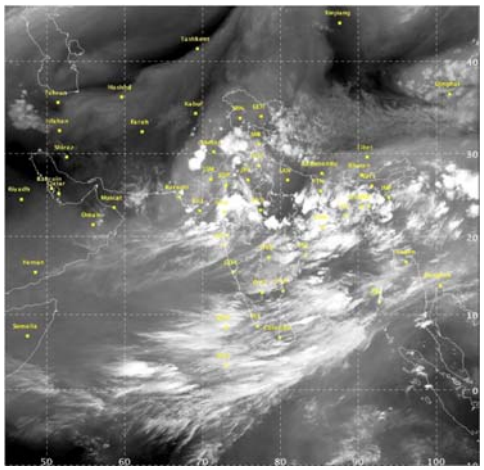


Figure 1. The original Satellite Water Vapour image.

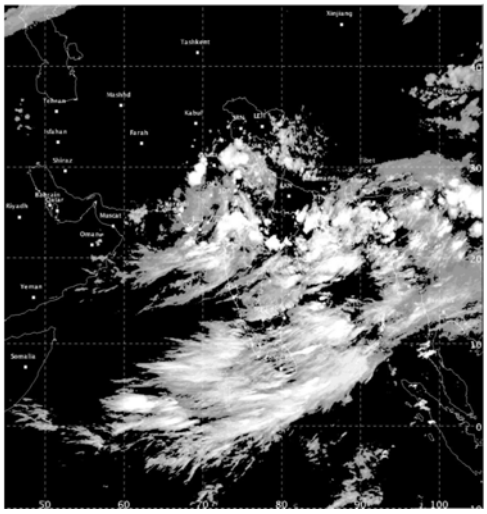


Figure 2. Segmented Image obtained after applying AABT algorithm

B. Region Seperation

In this step, the region separated out is the Western region containing Mumbai. This is done because we want to test our approach for Mumbai region only. The region is selected and cropped out to get a new image.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

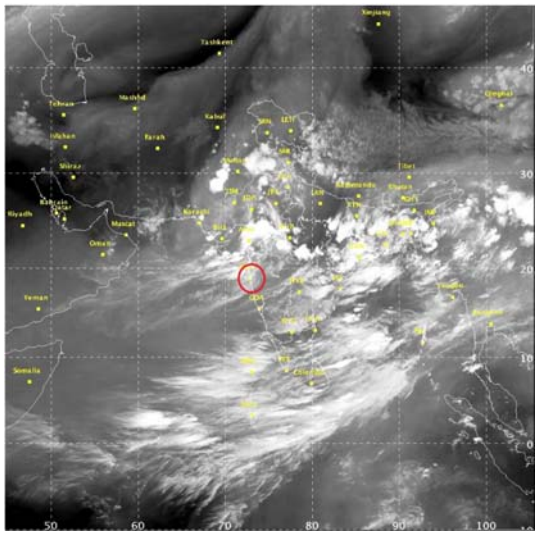


Figure 3. Region Selection for cropping.

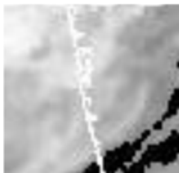


Figure 4. Final Cropped Image

C. Cloud Cover Extraction

The cloud cover percentage is calculated based on the equations given. It is calculated from the cropped image. For the above image the cloud cover percentage obtained is 96%.

VI. IMPLEMENTATION PROCESS FOR PREDICTION

After Image Processing and Segmentation is complete, the cloud cover information is extracted and is combined with the weather data table to create a consolidated database which is then used for predicting different parameters.

A. Dataset integration and preprocessing

The initial weather dataset consists of four weather parameters: Mean Temperature, Mean Humidity, Mean Wind Speed and Precipitation. The data derived from Image Processing and Segmentation consists of cloud cover over the chosen region. The values of cloud cover are divided into ten parts on the basis of 'greyness' of cloud cover, where a 'clear' sky is denoted by value '0' and an overcast sky is denoted by value '9'. The weather data is of Mumbai city from the period of June 2012 to June 2016 divided hourly. Therefore, there are 35,040 values in the dataset. The column for hourly cloud cover values is added the dataset

As the values in the dataset are of different ranges, the values are normalized before training them on the neural network. Normalization is done to scale down values of all columns of the dataset in the range of 0 to 1. The following is the normalized equation used.

$$norm_{data} = \frac{(data - \min(data))}{\max(data) - \min(data)} \quad (5)$$

Where data is the data matrix or the dataset.

B. Training non-target parameters using NAR Neural Network

We have assumed that precipitation is a phenomenon that depends on the remaining five columns, but it is still necessary to predict the other parameters. As the other parameters tend to exhibit a cyclic yearly pattern it is suitable to apply NAR model to predict the future value of data. Therefore, each column is initially independently trained except Mean Precipitation using the NAR Neural Network Model because it is the target. The training set is considered as the first 70% values of the dataset

This process is required as these predicted values for each of the remaining non-target columns are required as initial input for predicting the target value using the NARX model.

C. Training and Prediction of mean precipitation using NARX Neural Network

NARX model is trained with x[(t)] as all the input columns(i.e. Mean Temperature, Mean Humidity, Mean Wind Speed, Cloud Cover) and y[(t)] as the Mean Precipitation which is the target output. As explained above NARX model takes into account an external series which may affect the target series i.e. y[(t)].

The final prediction is to obtain precipitation values. These values can be easily obtained by testing the above trained NARX neural network using the input values obtained from training the individual columns through the NAR neural network as explained in the previous section.

D. Denormalization and Output

The data obtained from the output would still be in normalized form and it is necessary to de-normalize. The data is de-normalized using the following equation.

$$data = minVal + norm_{data}.* (maxVal - minVal) \quad (6)$$

Where minVal is the minimum value in the normalized data matrix, maxVal is the maximum value in the normalized data matrix and normdata is the normalized data matrix. The de-normalization of the data is performed column-wise.

VII. EVALUATION AND RESULTS

The model discussed above was created and implemented using the MATLAB's Artificial Neural Network Tool. Figure. shows a snapshot of the data used to train and test the model. Training, testing and cross-validation data were split in the ratio of 70:15:15 so as to obtain the best results. Many visualization methods were used such as, Performance Graph Plot and Error Histogram Plot to determine the quality of results. Mean Square Error (MSE) was one of the main factors used for determining the quality of the obtained results and is depicted as a part of the performance graph plot. Other parameters for testing include 3 layers of hidden neurons. Weights were assigned arbitrarily, i.e. according to the rules assigned to the toolbox.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I. A SAMPLE FROM THE DATASET

Mean Temperature(C)	*Mean Humidity (%)*	*Mean Wind speed(km/h)*	*Cloud Cover*	*Mean Precipitation(mm)*
28	77	11	6	13.97
29	75	10	5	0.25
29	75	10	4	0
31	71	13	3	0
28	79	11	5	12.95
29	76	13	2	0

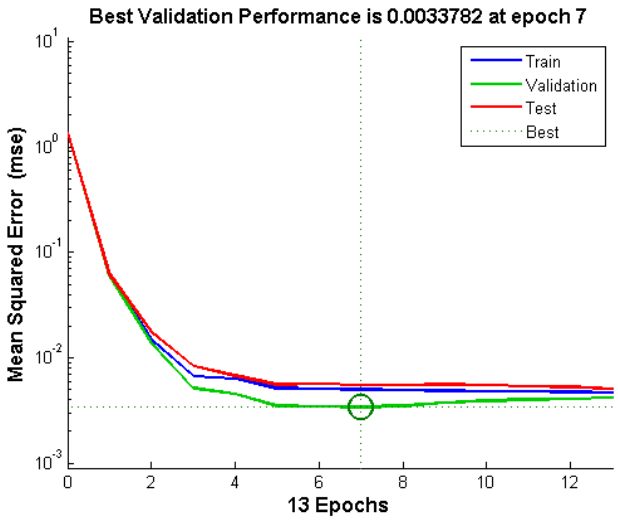


Figure 5. Performance Graph

The performance graph in Fig. 5 denotes the flow of the model during training of different aspects of the model i.e. training, validation and testing. The X-axis denotes mean square error and Y-axis denotes the epochs (a unit of time). The green circle denotes the point having minimum MSE after convergence. The convergence of the graph is a clear indication that there is no over fitting. The best performance of the model was at epoch 7 where the MSE was 0.0033782. This is the best performance point of the model.

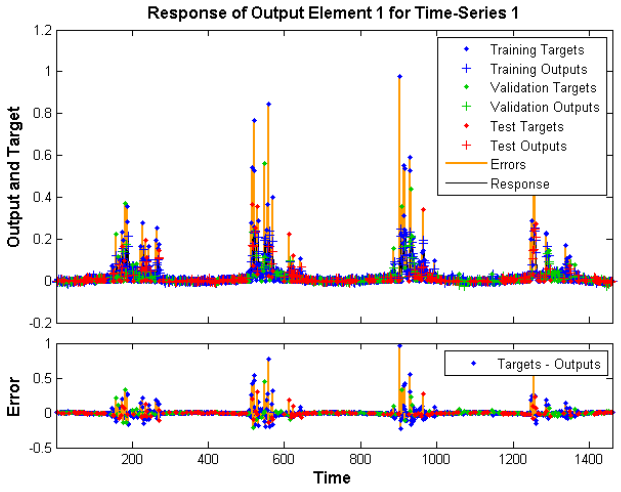


Figure 6. Time-Series Response Graph

The Time-Series Response Graph as shown below in Fig. 6 depicts the flow of error with respect to time. The lower graph indicates how much the model deviates from zero error over the period of training. The four major peaks in the graph indicate the arrival of the monsoon season in Mumbai for each year, during the four-year period.

VIII. CONCLUSION

For an automated weather satellite image interpretation system, one of the key steps is image segmentation. In this process, significant cloud features are extracted from the image and prepared for the next step in the process. AABT is designed to provide a fast and accurate method of image segmentation which is simple to implement as well. The segmentation results are provided quickly and with potentially enough accuracy to be integrated into a complete automated weather interpretation system or for cloud cover estimation.

Furthermore, in case of the neural network model it can be successfully concluded from the above results that this unique combination of NAR and NARX neural network produce a positive result and the prediction is accurate to a good extent although there is always a possibility of error as weather phenomenon are such that the features required to be incorporated to create an extremely efficient model are very high, varied and in many cases incalculable. Although efforts in this area will always develop the scientific community as well as the world.

REFERENCES

[1] Yo-Ping Huang , Tsun-Wei Chang and Li-Jen Kao, “Using Fuzzy SOM Strategy for Satellite Image Retrieval and Information Mining”, Systemics, Cybernetics And Informatics vol. 6 number 1 pp. 56-61.

[2] Craig M. Wittenbrink, Glen Langdon, Jr. , “Feature Extraction of Clouds From GOES Satellite Data for Integrated Model Measurement Visualization”, Technology & Engineering 2010

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[3] Du Huadong, Wang Yongqi, Chen Yaming, "Studies on Cloud Detection of Atmospheric Remote Sensing Image Using ICA Algorithm", 2009 IEEE.
[4] Chiang Wei , Wei-Chun Hung and Ke-Sheng Cheng, "A Multi-spectral Spatial Convolution Approach of Rainfall Forecasting Using Weather Satellite Imagery", Journal of Advances in Space Research, 2006
[5] "Localized Precipitation Forecasts from a Numerical Weather Prediction Model Using Artificial Neural Networks" by Robert J. Kuligowski and Ana P. Barros.
[6] M.W Gardner and S.R Dorling in their paper titled "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences".
[7] Isaac J.H. Leung, James E. Jordan, "Image Processing for Weather Satellite Cloud Segmentation", Canadian Conference on Electrical and Computer Engineering pp. 953-956.
[8] Journal of Signal and Information Processing Vol.5 No.2(2014), Article ID:45990,10 pages DOI:10.4236/jsip.2014.52007 A Nonlinear Autoregressive Approach to Statistical Prediction of Disturbance Storm Time Geomagnetic Fluctuations Using Solar Data.
[9] S. A. Billings. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, Wiley, ISBN 978-1-1199-4359-4, 2013.

Advanced Research in Electrical and Electronic Engineering
p-ISSN: 2349-5804; e-ISSN: 2349-5812 Volume 2, Issue 13 October-December, 2015, pp. 13-17

http://www.krishisanskriti.org/Publication.html

Cloud Image Analysis for Rainfall Prediction: A Survey

Minakshi Gogoi[1] and Gitanjali Devi[2]

[1]Dept. of Computer Science and Engineering Girijananda Chowdhury Institute of Management and Technology
[2]M.Tech., 3rdsem, Dept. of Computer Science and Engineering Girijananda Chowdhury Institute of Management and Technology
E-mail: [1]minakshi_cse@gimt-guwahati.ac.in, [2]gdevi43cse@gmail.com

Abstract—*The existence of clouds in the sky has a great impact, as it contains useful information for prediction of rainfall. Rain is the essential part of our ecosystem and it is responsible for most of the fresh water on the Earth. The rainfall plays a vital role for the balancing the heat ratio of Earth as well as water for hydroelectric power plants and crop irrigation. As almost 70% of population of any nation depends on agriculture, so rain can be the deciding factor of any nation's economic condition. Earlier prediction of rainfall was done by traditional means when technology was not developed. Later, with the development of various technologies prediction of rainfall became an easier and accurate way by studying different types of cloud image. Prediction of rainfall is possible with cloud images which may be either digital/satellite image. Prediction of rainfall can provide us with various pros and cons directly or indirectly hampering living beings on Earth. Early prediction of rainfall can provide us to preparedness of various disasters. But accurately forecasting of rainfall is complex as the clouds keep on changing. Clouds keep on changing depending upon different seasons. Prediction of rainfall can be done during the monsoon season where we may get the required result. As, previous work done, it is known that the rainfall clouds are the Nimbostratus and Cumulonimbus .Clouds like Cumulus can produce rain at very rare chance. Till now research is going onto predict rainfall using various technologies. But accurate forecasting of cloud image is a complex process. Image processing is one of the eras with its new technologies that can be used for detecting the early information about rainfall. In this paper a brief view of different types of clouds is discussed. Also, a brief analysis to highlight the possibilities of image processing methods is carried out that can be focused for effective prediction of rainfall.*

Keywords*: Cloud images, Cloud Types, Rainfall prediction, Image processing methods.*

1. INTRODUCTION

Clouds are important for balancing the Earth's climate, weather and temperature. A small change in the clouds could change the weather heavily. The presence of clouds in the sky can predict that there will be changes in the weather. The prediction of weather will be possible by understanding different type of clouds. Clouds are needed to be classified so that it become easier to detect which type of cloud it belongs to. So, that weather can be accurately predicted. Depending upon their height and characteristics, the clouds can be classified into various types. Clouds can be characterized based upon their shape, color, density, degree of cover, altitude at which they occur. There are three basic types of clouds and seven other types of clouds. The basic clouds are the Stratus, Cirrus, and Cumulus. Clouds can also be classified based on their altitude i.e., High Clouds are the "Cirrus", the Middle Clouds are the "Alto", and the Low Clouds are the "Stratus". It has been found that the rainfall clouds are the Nimbostratus and Cumulonimbus. Other clouds like Cumulus will produce rain at rare chances. Clouds can block the sunlight rays reaching the Earth's surface due to which the Earth's surface tends to be cooler. Clouds have different shapes and structure due to which prediction of weather become complex.

Rainfall forecasting is important for agriculture and living things. Rainfall being an important part of agriculture helps in productivity of various vegetables, fruit, flowers etc. Some of the products of flower, crop, and vegetables are exported. Crops like rice, corn are grown in heavy rainfall area. Rainfall prediction is needed so that no damages occur to living beings, crops, land etc. If prediction is done accurately it can save many lives which may occur due to floods. Rainfall prediction is necessary to determine how much of the moisture is available for agriculture and how much of them have been run off to the rivers and streams. However, in excess rainfall may cause damages to agriculture moisture which may be carried out by the rainfall to the streams or rivers. As a result of which floods from the rivers and streams may cause damages to roads, and loss of various sediment and chemicals. Observation of rainfall is needed to know the amount of moisture present in soil, drought and flood conditions. Effective and accurate prediction of rainfall is necessary to prepare for floods and agriculture. Traditionally, various measures have been used to predict rainfall. Prediction of rainfall becomes a harder because the status of cloud/sky keeps on changing. The images taken for prediction may be digital/satellite images. These images are taken and types of cloud/sky are found. Previously many techniques of image processing have been carried out successfully. Image

processing techniques can be used to measure the cloud, sky, rain status. Types of cloud can be found by using various technique of image processing. Image processing is used to enhance various features of the images taken to predict rainfall. Once type of cloud is known it becomes easier to predict rainfall. To accurately predict rainfall, the cloud status and sky status must be known.

Clouds are the source for prediction of rainfall. If clouds are present in the sky, chance of rainfall is more. We cannot predict rainfall just on seeing the clouds. In this paper, reviews about the clouds for prediction of rainfall have been focused and various image processing methodologies are discussed.

2. RAINFALL PREDICTION

Rainfall prediction helps us in various ways if it is forecast accurately. Prediction becomes possible by studying various types of clouds by knowing their status and types which depends on height, altitude, density, color. Observations of rainfall are necessary in various fields and for various purposes. Though prediction of rainfall provides us useful measures, but they are hard to predict accurately

.Advantages: Meteorology focuses on weather processes and forecasting. Meteorology department is useful in giving information of floods, agriculture and prediction of rainfall. Rainfall prediction is useful for agriculture purposes to observe soil moisture, effectiveness for applying fertilizers, pesticides and herbicides to crops. Accurately forecasting of heavy rainfall can allow for warning and preparedness of floods. Rainfall can monitor and forecast drought, river stages, and water quality [1].

Disadvantages: Prediction of rainfall is not accurate as clouds are very complex and keeps on changing, which can cause streams and rivers to overflow. Overflow of rain water can cause severe damages to human beings, transport and loss of sediments and chemicals causing loss to country's economic growth [1].

3. ISSUES IN DESIGNING A RAINFALL PREDICTION ANALYSIS

The proposed system design steps are as follow:

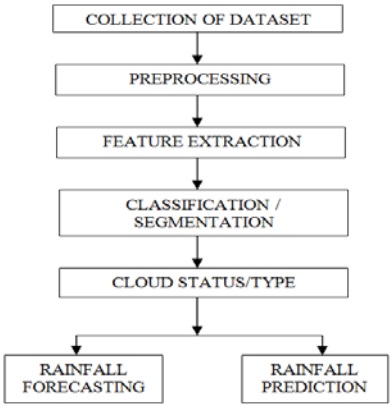


Fig. 1: Proposed system design

a) In the first step, the cloud images are collected and stored (digital/satellite).

b) The image stored will be preprocessed. Preprocessing is the step where noise filtering, sharpening, contrasts stretching, histogram modification and correction of distortion may be applied.

c) Next, feature extraction method is applied to an image. Feature extraction involves extraction of shape, color or texture of an image.

d) Segmentation methods can be applied to various cloud images to segment some of properties of the cloud into group of pixels which are homogeneous with respect to some criterion.

e) Cloud status/type are used with the image data to forecast rainfall

f) After combining all, the data are inputted and find rainfall prediction.

4. TYPES OF CLOUD IMAGE

Table 1: Basic cloud types and description [14]

Cloud type	Description
Cirrus	Thin, white and feathery appearance and mostly white patches or narrow bands.
Cirrocumulus	Thin white bands or ripples, sheet, or layered of clouds without shading.
Cirrostratus	High, milky white like appearance. They are transparent, whitish veil clouds with a fibrous (hair-like) or smooth appearance
Altocumulus	Bumpy rounded masses, cotton ball appearance, white and/or gray patch sheet or layered clouds.
Altostratus	Transparent blue/gray clouds sheets or fibrous clouds that totally or partially cover the sky.
Nimbostratus	They are continuous rain cloud also known as storm cloud.
Stratocumulus	Gray or whitish layer with sheet, or layered clouds which almost always are dark.
Stratus	Cover large portion of sky, thin, sheet-like, gray and thick.
Cumulus	Cauliflower like appearance with bulging upper parts.
Cumulonimbus	The thunderstorm cloud, heavy and dense cloud in the form of a mountain or huge tower.

Clouds are classified by their shape, temperature, color, density, spectral clustering analysis, training of rule based systems or neural networks and image processing techniques. Classification of clouds are highly time dependent because temperature, type of cloud dependence changes in different latitudes through different seasons. Different types of cloud are present out of which three basic cloud forms are the Cirrus, Stratus, and Cumulus. These forms are further refined into 10 other types based on their height and texture as shown in Table 1.

High Level Clouds (above 20, 000 feet) are the Cirrus, Cirrocumulus, Cirrostratus, and Cumulonimbus.

Middle Level Clouds (between 6500–20,000 feet) are the Altocumulus, Altostratus.

Low Level Clouds (below 6500) are the Nimbostratus, Stratocumulus, Stratus, and the Cumulus.

4.1 Prior of related work

K.Kaviarasu et.al [1] describes some novel methods like K-means Clustering and Wavelet using digital cloud image. The status of sky was found using Wavelet. The status of cloud is found using Cloud Mask Algorithm and histogram equalization. The type of cloud can be found using K-means Clustering.

In 2015 Niyati Salot et.al [3] has studied various techniques of image processing to predict rainfall. Various methodologies/techniques of image processing have been discussed in processing of the dataset.

Dr. Mohamed Mansoor Roomi et.al [4] in 2012 explained an automatic detection of weather forecasting by segmenting the various cloud images, which do not depend upon translation, rotation and scale. SURF (Speeded up Robust Features) is used here for feature extraction. The extracted images are then segmented using the Otsu thresholding.

Further Dimple Jayaswal et.al [5] proposed a way for retrieving satellite images. It first works on feature extraction of the image and then compares the image with the images in the database. The image retrieved process contains the satellite image where it is pre-processed, feature is extracted and calculation is done. Interpretation and evaluation is done and the final image is retrieved.

It has been found that in 2006 D.K.Richards and G.D.Sullivan [6] used the feature extraction method of color and texture to the cloud and sky images using digital image. A single method alone cannot distinguish different types of cloud. By using Bayesian scheme the classification of various features can be improved.

Again in 2015 Dimple Jayaswal et.al [7] gives a review of various image retrieval image techniques which can be applied for satellite images. In feature extraction the shape, color, textures are extracted. Content-based image retrieval (CBIR) method is used for image retrieval.

Again, some ways of image retrieving techniques was focused by Neha Jain et.a l [8] in 2013. A brief study of the techniques plus its various advantages and disadvantages has been discussed here.

Yanling Hao et.al [15] said that cloud images are useful image which contains a lot of information. To acquire this information various image processing methods and feature extraction method are used. CBIPIR (Content Based Cloud Image Processing and Information Retrieval) is an important retrieval technique in image processing. Features like shape, color, texture, edge are extracted from the cloud images and stored in database for further analysis.

5. LEVEL OF PREDICTION

There are various ways to predict rainfall. Long before when technology was not developed to predict the weather, people relied on observation, patterns. Once these methods are practice and become attuned to the sky, the air, and animal behaviors, it's possible to predict the weather quite reliably .We have many traditional ways to predict rainfall like the presence of red sky, observing the sky and many more. But, later with the development of technologies prediction of rainfall became an easier and accurate. We have various methods like

a) Traditional method: These methods used clouds to predict rainfall. By looking at the various patterns of clouds rainfall prediction was done. The other ways were like the presence of red sky, the existence of clouds, look for rainbow in the west, and the direction of flow of winds. These were the methods in where no techniques were applied. By natural means the prediction of rainfall was done.

b) Statistical method: These methods are based on collection of data for monthly values or yearly values of rainfall. These can be aggregated into region wise, district wise, state wise or country wise. The climate conditions have been measured adequately for many years, making it possible to define what is “normal” and what an “extreme” is. The rainfall data is collected over a period through which estimation of heavy rainfall can be got. The data are compared over the previous stored or collected information of rainfall.

c) Numerical method: Numerical weather prediction uses mathematical models of the atmosphere and oceans to predict the weather , based on current weather conditions. A number of global and regional forecast models are run in different countries worldwide, using current weather observations relayed from weather satellites and other observing systems as inputs.

Mathematical models based on the same physical principles can be used to generate either short-term weather forecasts or longer-term climate predictions; the latter are widely applied for understanding and projecting climate change. But, predicting rainfall by manipulating the dataset and performing complex calculation requires powerful supercomputers in the world. The forecast still only extends for six days. The accuracy of numerical prediction can be affected by density and quality of input.

5.1. Cloud images using image processing methods

Work has been carried out for various cloud images to predict rainfall using various techniques. Cloud images can be either digital or satellite images. The various image processing methods which can be applied are: Image storage and

manipulation, Image enhancement, Image restoration, Image analysis, and Image reconstruction. Cloud based image processing techniques which were carried earlier for cloud images are [5, 8], and [9].

5.1.1. Feature extraction

It is one of the important methods of image processing. Various feature extraction method are present but choosing the right image features for any system is important because it may affect every aspect of a retrieval system. Low level features are explored as they can be computed automatically. We discuss below some of the characteristic of the features.

a) Shape based feature: To retrieve image by shape is the most obvious requirement at the starting level. Natural objects are recognized by their shape mainly. For various stored image the feature characteristic of object shape are computed for the objects in the image. When queries are given for stored images, the images are retrieved by computing the same set of features whose features closely match to that of the query. Two main types of shape feature are used commonly, the global features (aspect ratio, circularity and moment invariants) and the local features (like sets of consecutive boundary segments) [8].Shape matching of three-dimensional objects is a more challenging task.

b) Color based feature: Here for stored image in database a color histogram is computed which shows the proportion of clouds of each pixels in the image. The color histogram for every image is then stored in the database. When querying for the image the user can either give the proportion of each color, or give an example image from where a color histogram is calculated.

c) Texture based feature: Texture based features may not be very useful. But to match texture similarity can be useful to distinguish between different areas of images with similar color. Different techniques are present to brightness of the selected pair of pixels from each image can be calculated. From these the texture of various images can be calculated.

5.1.2. Classification methods

Two learning methods are compared- supervised and unsupervised. The main difference between supervised and unsupervised classification is the use of collection of training dataset images. In supervised classification a collection of images are taken. A single image is compared with a number of images to get the required output. But, in unsupervised classification collection of training dataset is not available. The output result is based on software analysis of images without sampled images.

6. CLOUD IMAGE DATABASE

Cloud image collected may be digital/satellite images. In case of digital images the image are collected from ground level through digital camera or through web. While, satellite images are collected either from meteorology department or collected on daily basis from some meteorology websites. The images are stored and manipulated in different formats in the database. Images can be stored in different formats which may be Binary, TIFF, JPEG, GIF, PNG or BMP image. After storing the image in the database, different functions can be applied to read the images from the database.

7. EVALUATION OF PREDICTION ACCURACY

Prediction of rainfall has been carried out for more than 20 years. Weather forecasts have improved over the last 20 years. The three-day forecast delivered today is much better than that that of one-day 20 years ago. There are various methods to predict rainfall. But, accurately predicting rainfall is a complex process. The monsoon forecast can be closer to perfection and rains can be predicted more accurately. Prediction of accurate monsoon is important for an individual farmer as well as for the government to take policy decisions. However, forecast is always unpredictable.

Earlier traditional methods were used which gave results. The methods were carried on natural means. Earth's atmosphere is not as it was 20 years back. Due to changes in the Earth's atmosphere by various sources prediction through traditional method does not play well.

Statistical method required data over a long period of time. The data collected over years are compared to the current situation. The climate changes and does not remain constant. So, prediction of rainfall through this method cannot be dependent totally.

Prediction of rainfall through cloud images are growing fast and works are carried by researcher's to give accurate results. Cloud image may be digital or satellite images. Digital image are easily available but resolution of image is less. To predict rainfall the clouds should be clear in the sky but digital image may lack this property. Prediction of rainfall done by digital cloud image may not be clear in structure which may result in poor prediction of rainfall. Prediction of rainfall through satellite image can be accurate. Since satellite image has clear cloud structure. Image processing method for cloud image can be carried out to evaluate the prediction of rainfall accurately. Methods of image processing can be applied to various cloud images to study the status and type of clouds. Works have been carried out in this field and further research can also be done to get accurate results.

8. CONCLUSION & FUTURE OF RAINFALL PREDICTION ANALYSIS

This paper gives a review of different cloud image used for prediction of rainfall. Some of the researchers used satellite images of cloud and sky and some of them used digital image which is a cheaper one. Through the above article, an idea of the direction of research regarding to the types and status of cloud are studied for detecting the early information of rain.

This study shows that, the instantaneous research on rainfall prediction of cloud image as compared to others is found to be less. It needs further research on this line.

REFERENCES

[1] K.Kaviarasu, P.Sujith and Mr. G. Ayaappan, "Prediction of Rainfall using Image Processing",2010 IEEE International Conference on Computational Intelligence and Computing Research, ISBN : 9788183713627.

[2] Wei Shangguan, Yanling Hao, Zhizhong Lu and Peng Wu, "The Research of Satellite Cloud Image Recognition Base on Vibrational Method and Texture Feature Analysis", Industrial Electronics and Applications, ICIEA 2007, 2nd IEEE Conference, Volume, Issue, 23-25 May 2007.

[3] Niyati Salot and Dr.Priya R.Swaminarayan, "A Survey on Rainfall Forecasting using Image Processing Technique", IJITE Vol.03 Issue-02, (February, 2015), ISSN: 2321-1776.

[4] Dr.Mohamed Mansoor Roomi, R.Bhargavi and T.M.Hajira Rahima Banu, " Automatic Identofication of Cloud Cover Regions using SURF", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2,April 2012, DOI : 10.5121 / ijcseit . 2012. 2214.

[5] Dimple Jayaswal, Vishal Shrivastava, "A Literature Review on Satellite Image Retrieval Techniques", International Journal of Computer Science Trends and Technology (IJCST) - Volume 3 Issue 2, Mar-Apr 2015, ISSN: 2347-8578.

[6] K. Richards and G.D. Sullivan," Estimation of Cloud Cover using Colour and Texture", BMVC 1992, doi:10.5244/C.6.45

[7] Dimple Jayaswal, Vishal Shrivastava, "A Proposed System for Satellite Image Retrieval", International Journal of Engineering Trends and Applications (IJETA)–Volume 2 Issue 2, Mar-Apr 2015, ISSN: 2393–9516."

[8] Neha Jain, Sumit Sharma and Ravi Mohan Sairam, "Content Base Image Retrieval using Combination of Color, Shape and Texture Features", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume-3 Issue-8 March-2013.

[9] Jay Narayan Thakre and Divakar Singh, "A Survey on Knowledge Discovery from the Satellite Image Using Association Rule Mining ", Volume 3, Issue 8, August 2013, ISSN: 2277 128X.

[10] Noureldin Laban, Ayman Nasr and Motaz ElSaban Hoda Onsi : "Spatial Cloud Detection and Retrieval System for Satellite Images ",(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 12, 2012.

[11] B.Ramesh, Dr. J. Satheesh Kumar, "Cloud Detection and Removal Algorithm Based on Mean and Hybrid Methods ", International Journal of Computing Algorithm, ISSN: 2278-2397, Vol 02, Issue 01, June 2013.

[12] D. L. Shrestha, D. E. Robertson, Q. J. Wang, T. C. Pagano, and H. A. P. Hapuarachchi," Evaluation of Numerical Weather Prediction Model Precipitation Forecasts for Short-term Streamflow Forecasting Purpose", www.hydrol-earth-syst-sci.net/17/1913/2013, doi:10.5194/hess-17-1913-2013.

[13] Harinder Kaur Narain, Neelofar Sohi, "Review: Segmentation Algorithms for Cloud Detection", http: // www.ijesrt.com, ISSN: 2277-9655.

[14] Craig M. Wittenbrink, Glen Langdon, Jr., and Gabriel Fernandez," Feature extraction of Clouds from GOES Satellite Data for Integrated Model Measurement Visualization".

[15] Yanling Hao, Wei ShangGuan, Yi Zhu, and YanHong Tang," Contented-Based Satellite Cloud Image Processing and Information Retrieval", Springer-Verlag Berlin Heidelberg 2007, LSMS 2oo7, LNCS 4688, pp.767-776.

[16] R. Samuel Selvaraj and Raajalakshmi Aditya," Statistical Method of Predicting the Northeast Rainfall of Tamil Nadu", Universal Journal of Environmental Research and Technology, Volume 1, Issue 4: 557-559, ISSN 2249 0256.

Weather Forecasting Using Digital Image Processing

Mrs. G. Gowri Pushpa
Department of Computer Science and Engineering
ANITS, Vishakhapatnam, Andhra Pradesh, India
gowripushpa11@anits.edu.in

Hemaja Patoju
Department of Computer Science and Engineering
ANITS, Vishakhapatnam, Andhra Pradesh, India
phemaja@gmail.com

G. Sai Charan
Department of Computer Science and Engineering
ANITS, Vishakhapatnam, Andhra Pradesh, India
charan4899@gmail.com

M. Kali Charan
Department of Computer Science and Engineering
ANITS, Vishakhapatnam, Andhra Pradesh, India
Marpukalicharan@gmail.com

Pranita Jagtap
Department of Computer Science and Engineering
ANITS, Vishakhapatnam, Andhra Pradesh, India
2494pranita@gmail.com

Abstract- To predict the conditions of the atmosphere for a given location Weather Forecasting is used. It is the application of science and technology. Weather forecast is more helpful for people as it predicts how the future weather is going to be and people may plan accordingly. Farmers will be the most beneficial one's as they may know the rainfall prediction and grow crops accordingly. The weather forecast can be done in many ways like using the previous data or analyzing the current clouds. The authors predict the weather using the status of the clouds. The author used methodologies like Normalization, Clustering, and Cloud mask algorithm to predict the weather more accurately. Normalization is done using RGB values of each pixel. In many fields of research and in industrial and military applications Digital-image processing has become economical.

Keywords - Weather forecasting, normalization, clustering, and cloud mask algorithm.

I. INTRODUCTION

Weather forecasting [1] means predicting the weather and telling how the weather changes with change in time. Change in weather occurs due to movement or transfer of energy. Many meteorological patterns and features like anticyclones, depressions, thunderstorms, hurricanes and tornadoes occur due to the physical transfer of heat and moisture by convective processes. Clouds are formed by evaporation of water vapor. As the water cycle keeps on evolving the water content in the clouds increases which in turn leads to precipitation. This is how the convective process happens and also the change in weather. Many factors like temperature, rainfall, pressure, humidity, sunshine, wind and cloudiness are considered for predicting the weather. It is also possible to identify the different types of clouds associated with different patterns of weather. These patterns of weather help in predicting the weather forecast.

In the past, people used barometric pressure, current weather conditions, sky condition to predict whereas now there are many computer based models that consider the atmospheric factors to predict the weather. These methods are not accurate and the reason is due to the chaotic nature of the atmosphere as it keeps on changing. Even predicting weather for a longer period of time will not be accurate that is why most of the current forecasting [1] models predict weather

only for a couple of days not more than 10. The accuracy gets reduced with increase in time. We researched some of the papers and cloud types for this paper. They are

Machine learning applied to weather forecasting:

In this paper, details of weather for the past 2 days are considered. Those details are considered as input and performing linear regression and variation of functional regression, output is obtained. The output is weather for next 10 days. Generally the classification of weather gives 9 classes: clear, scattered clouds, partly cloudy, snow, thunder storm, rain, overcast, fog, mostly cloudy [2]. The dataset considered classified all those into 3 classes: moderate cloudy, very cloudy, precipitation. The least mean square error for the linear regression and variation on functional regression is calculated and learning curves are drawn in this paper. Linear regression is low biased with high variance model whereas functional is exactly opposite to it. Collection of more data can improve the linear regression model [2]. Hence the author suggests considering 4 to 5 days of data as input to the model.

Analysis on various techniques for weather forecasting:

1) Support Vector Machines: To predict the maximum temperature of a required location Support Vector Regression (SVR) is used. It performs better than MLP which is trained with back propagation algorithms as it minimizes the upper bound on generalization error. By selecting proper parameters it can replace neural networks based models for applications of weather prediction.

2) Time Series Analysis for Weather Forecasting: Data groups and data variables in the specified time are captured by Time Series Analysis. By comparing actual and predicted values of temperature, the forecasting reliability was evaluated. The results show that important tool for temperature forecasting is network.

3) Prediction of Weather by using Back Propagation Algorithm: Wind, humidity, rainfall and temperature are the parameters recorded using sensors. Using these sensors weather forecasting and processing information is transferred [3]. It classifies compares and predicts the change in other weather parameters by changing any one parameter value that those sensors recorded. A 3 layered neural networks trained with the existing dataset to develop a relation among the parameters of weather that are non-linear.

4) Fuzzy Logic Based Rainfall Prediction model: Two components are made in a developed fuzzy logic model where one is knowledge based and the other is fuzzy reasoning or decision making. Using fuzzification and defuzzification operations outputs are predicted compared with actual rainfall data [3]. A fuzzy model that is well developed is capable of handling the data that is scattered and shows flexibility in modelling weak input and output variable relationship.

Weather forecasting using data mining research based on cloud computing:

A modern method is developed which is service oriented architecture for the weather information systems that forecasts weather using data mining techniques. The method uses Artificial Neural Network and Decision tree Algorithms and meteorological data collected in specific time [4]. It presents the best results for generating classification rules for the mean weather variables. The model predicts temperature, rainfall and wind speed. Cloud computing reduces the cost of infrastructure and storage as it ensures secure reliable and efficient services for the user.

Cloud image analysis and classification:

Generally clouds are classified on the parameters like temperature, shape, colour, density, spectral clustering analysis. Cirrus, Stratus and Cumulus are the 3 basic types of clouds. Based on their height and texture these clouds are divided into 10 other types [5].

1. Cirrus
2. Cirrocumulus
3. Cirrostratus
4. Altocumulus
5. Altostratus
6. Nimbostratus
7. Stratocumulus
8. Stratus
9. Cumulus
10. Cumulonimbus

The dataset we considered is named “HYTA”. It consists of various images of all types of clouds. We considered 4 clusters for all the types of clouds namely: clear sky, sunny, cloudy and sunny, rainy. For each type of cluster this HYTA dataset consists of nearly 8 to 10 images. Every image consists of only plain sky with respective clouds and no other objects like buildings, trees and poles. In some images the sun might appear along with the sky and clouds. Along with this standard dataset we considered 4 different datasets. Every data set consists of more than 10 photos for each type of cloud. These datasets are considered to compare the outputs obtained and check the accuracy of the model developed. This comparison will be useful for the future development of the model.

II. PROPOSED ALGORITHM

Weather forecasting can also be done by using satellite images but acquiring the satellite images is more difficult and would even cost high. Even predicting using the satellite images needs more technology. So, we are using digital image processing techniques which process the images of the sky like normalization, cloud masking algorithm and k-mean algorithm.

Every system should be divided into modules for better understanding and execution. Dividing into modules helps the programmer and client to work and use the system efficiently, respectively. If any system is not divided into modules and worked as a whole, then there came a numerous errors. Even we find difficulty in correcting those errors. It is must and should to divide the total project into modules and work on each and every module independently to get effective results. Our total system is divided into three modules namely:

1. Normalization of Image
2. Cloud masking algorithm
3. K-means clustering.

2.1. Normalization of Image:

Step 1: Pixel values for each and every pixel are considered. Pixel value consists of red, blue and green colour's values. These values are extracted from the image with the help of pre-defined libraries in python.
Step 2: Now with the help of these pixels, we must change the intensity range of the pixels to [0,1] and increase the intensity to get a clear distinction between the clouds and the sky. Hence the digital picture is normalized [6].
Step 3: The input image can be of any digital image with the extension .jpg, .jpeg, .png.
Step 4: The output of this module would be a normalized image of the given digital image which seems likely to be a black and white or gray scale image.

Figure 1. Original image of cloud

Figure 2. Image after performing normalization

2.2. Cloud masking Algorithm:

Step 1: After normalization, a mean value is generated by adding all the pixel values and by dividing it by the total no of pixels. With the help of this mean value we differentiate the clouds from the input image.
Step 2: Now we have to extract the feature of the cloud part by again finding the mean value of the cloud area which will be used a feature in the next process [7].
Step 3: This process is done for all the images in the dataset so that we get features of all images which will be used to cluster the images into groups. The output of this process is shown in Figure 3.

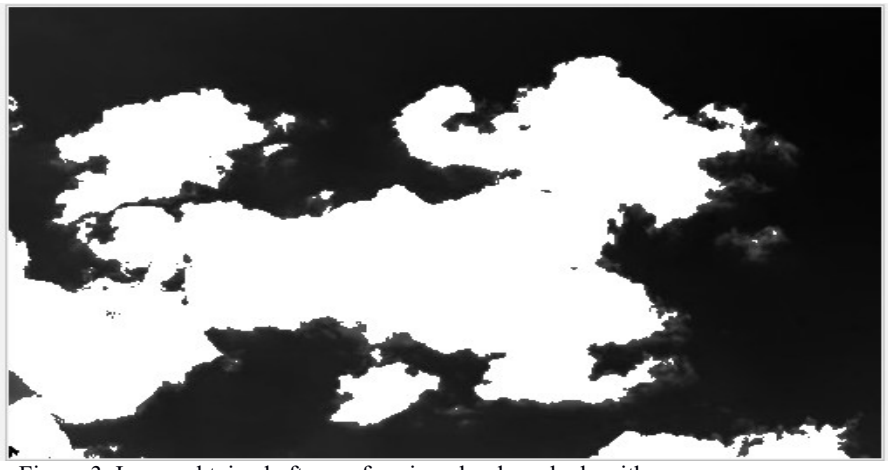


Figure 3. Image obtained after performing cloud mask algorithm

2.3. K-means Clustering Algorithm:

We considered clustering because for classification there would be less no of classes. But we considered ten types and hence we considered clustering rather than classification [8]. Here we considered the clusters of the clouds as we would divide the image based on the cloud mean point.

In this the clouds are divided into 10 clusters. The clusters are:

1. Cirrostratus
2. Cirrus
3. Cirrocumulus
4. Altocumulus
5. Altostratus
6. Stratus
7. Stratocumulus
8. Nimbostratus
9. Cumulonimbus
10. Cumulus

Among these the cirrostratus, altostratus, cirrocumulus, cirrus denotes sunny day. Nimbostratus and Cumulonimbus denotes rainy day. And cumulus, stratus, altocumulus and stratocumulus denotes cloudy day.

All the classification can be done depending on the mean threshold value. From the dataset after applying normalization and cloud masking algorithm we can get a threshold value for each and every cloud cluster. Based on that value i.e. the threshold value, the input cloud image is classified into a cluster [9]. Then based on the cluster we can forecast the weather.

ALGORITHM:

1. First we initialize k points, called means, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

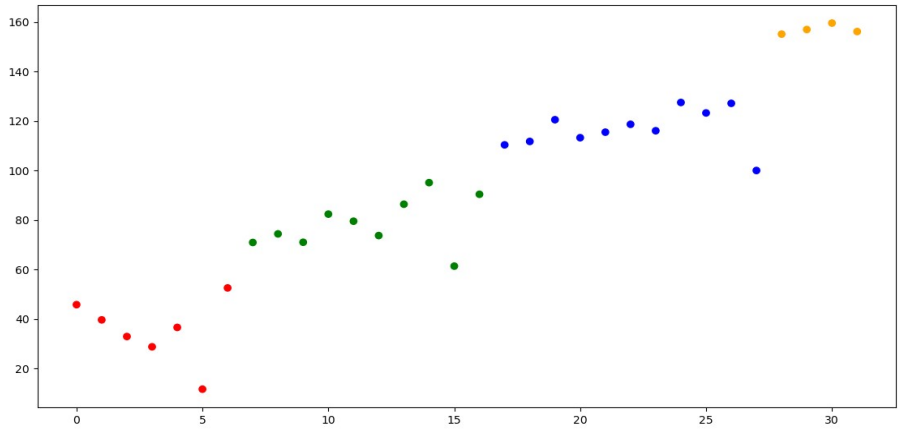


Figure 4. Scatterplot [10] graph showing different clusters formed after training the dataset

III. EXPERIMENT AND RESULTS

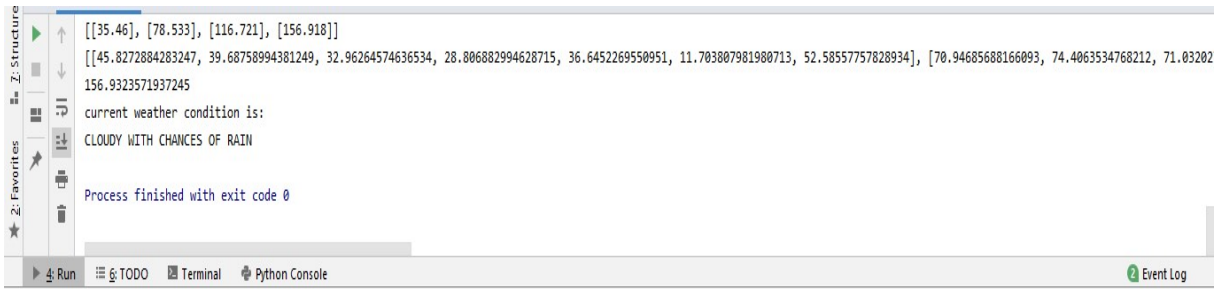


Figure 5. The generated output in the form of text for the given input image

The first line in figure 5 shows four different centroids of the obtained randomly from the HYTA dataset. And the second line shows 4 different lists. Each list contains the points generated after performing cloud masking algorithm to each image in the dataset. Every list consists of the points that are corresponded to that respective centroid. The third line is the mean point calculated out of the k means clustering algorithm. At last, it shows the result of the input image i.e. Figure 1 in the text format.

IV.CONCLUSION

Generally, any weather forecasting applications and sources would give the weather report of a particular area with the help of GPS [11] or using satellite information. Our model can give the weather condition at any point of time for any place with the help of the current cloud image at that place. In future this model can be developed as to predict the weather for the next few hours based on the image with the help of cloud analysis.

This paper can be extended to get weather forecast for the next few days; we can modify our system by using different algorithms and use that as an extension to our current project.

REFERENCES

[1] Weather Forecasting using Satellite Image Processing and Artificial Neural Networks by Nilay S. Kapadia , Urmil Parikh in IJCSIS vol 14,No.11, Nov 2016.

[2] Machine Learning Applied to Weather Forecasting Mark Holmstrom, Dylan Liu, Christopher Vo Stanford University, December 15, 2016.

[3] ANALYSIS ON THE WEATHER FORECASTING AND TECHNIQUES Janani.B, Priyanka Sebastian, Jan 2014.

[4] The Weather Forecast Using Data Mining Research Based on Cloud Computing. ZhanJie Wang and A. B. M. Mazharul Mujib,2017.

[5] Cloud image analysis for rainfall prediction: A survey by Minakshi Gogoi and Gitanjali Devi, Advanced Research in EEE, Oct-Dec 2015.

[6] https://www.mathworks.com/matlabcentral/answers/422493-how-to-do-normalized-blue-red-ratio-operation-of-an-image

[7] https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[8] https://www.geeksforgeeks.org/k-means-clustering-introduction/

[9] https://scied.ucar.edu/learning-zone/clouds/how-clouds-form

[10] https://www.science-emergence.com/Articles/How-to-create-a-scatter-plot-with-several-colors-in-matplotlib-/

[11] https://www.universityworldnews.com/post.php?story=20120620192857538