Snowfall Data Analysis and Visualization

Comparative Analysis of the Historical Snowfall Patterns using Big Data and ArcGIS

Group Name : Earth.ai

Puneet Singhania
Department of Computer Science and Engineering,
University of California, Riverside,
Riverside, CA, USA
psing088@ucr.edu

SID: 862327375

Rigved Patil
Department Of Computer Science and Engineering,
University of California, Riverside,
Riverside, CA, USA
rpati013@ucr.edu

SID: 862322104

Ritesh Singh
Department of Computer Science and Engineering,
University of California, Riverside
Riverside, CA, USA
rsing116@ucr.edu

SID: 862395274

Sanjana Senthilkumar
Department of Computer Science and Engineering,
University of California, Riverside
Riverside CA USA
ssent013@ucr.edu

SID: 862384460

Shadhrush Swaroop
Department of Computer Science and Engineering,
University of California, Riverside
Riverside CA USA
sswar010@ucr.edu

SID: 862394040

1. ABSTRACT

As we know, a huge amount of weather data is being generated by multiple weather stations and reported to the government datasets. These datasets are ever-increasing. In this project, we aim to do a comparative analysis of the historical patterns of snowfall data using big data technologies and visualize it using ArcGIS. This will help to get meaningful insights from this massive data and predict the snowfall pattern for which we use ML algorithms. As we know, snowfall and snow storms can cause loss of property and life. Historically, they have been known to cause multiple accidents, blizzards, disruption, and disturbances in healthcare. This makes it critical that we find efficient ways to obtain meaningful information. An analytical comprehension of historical snowfall patterns can help in improving preparedness and help understand the rate of change of snowfall. From our analysis, we were able to visualize difference in snowfall pattern year-on-year and able to identify that the change in temperature is an important factor for this.

2. INTRODUCTION

Every year, snowfalls and snowstorms cost millions when it comes to loss of life and property damage. This causes wide-scale disruption and disturbances. In order to avoid this, we need to plan preventive measures. Analyzing weather data collected by the weather stations around the globe on a regular basis can help give an idea about regions more perceptible to

extreme weather events. This data is present in enormous amounts and it is difficult to analyze and extract meaningful information from it using the traditional data analysis techniques. We strongly believe that, by using modern visualization techniques we can project data in easily interpretive ways. We visualize areas of high snowfall based on precipitation and temperature values. We were able to predict for the year 2019 using data from the previous twenty years. We provide a detailed breakdown of how we have come up with our solution and what different techniques/ methodologies we have used.

The overview of the sections provided below are as follows: **Section 3** explains the related works and literature survey done with respect to our project and their relevance to our project. **Section 4** provides an outline of the list of important components being used in our project. We have divided the components into several subcomponents like data processing, prediction and analysis and we provide a brief overview about the different parts of our web application. **Section 5** provides an explanation about the evaluation methodology and the parameters being used for the same. **Section 6** concludes the project report and discusses the solution proposed, summary of the project and the optimal results obtained. We also discuss potential areas we would like to explore in the future. **Section 7** provides the details about all the authors' contributions in a tabular format.

3. LITERATURE SURVEY

1. The ideology and factors influencing snowfall and its patterns

Description: This category provides papers that give insight into the correlation between snowfall and accidents, the influence of climate change on the intensity of snowfall, and parameters suitable to find hotspots and patterns of snowfall.

2. Snowfall data set

Description: Preprocessing Spatial Big Data is necessary in order to analyze it. Pre-processing is needed to get the data you want. Unordered, dirty, and unwanted data need to be removed from the raw data before it can be worked on. The missing values and noise are handled by cleaning the dataset. This category provides works related to genuine data sets that can provide inspiration for our project.

3. Snowfall prediction models and algorithms

Description: This category provides works related to the analysis of historical and current data and the utilization of different algorithms to process the data and provide meaningful insights. A comparative analysis of previous approaches can be done to find out the optimal approach and several parameters used in previous works can be integrated into our current project to boost its performance, efficiency, precision, and predictability.

4. Evaluation of ML models

Description: This category mainly consists of works that provide insights and parameters related to the evaluation of the ML models being used in relation to snowfall prediction. By doing such an evaluation, we can see in which type of region our predictions are more accurate and can provide metrics for the same.

5. Visualization

Description: Visualization is a technique where we are representing data in an understandable format. We are studying the approaches used in the past and how they can be used to enhance the visualization techniques in order to make them compatible with our project. We are currently working on/testing our model's output (test output) in order to check if this visualization report can be easily read and understood by everyday users. So that they too can benefit from it and ultimately help in their own research.

RELATIONSHIP DIAGRAM BETWEEN THE CATEGORIES

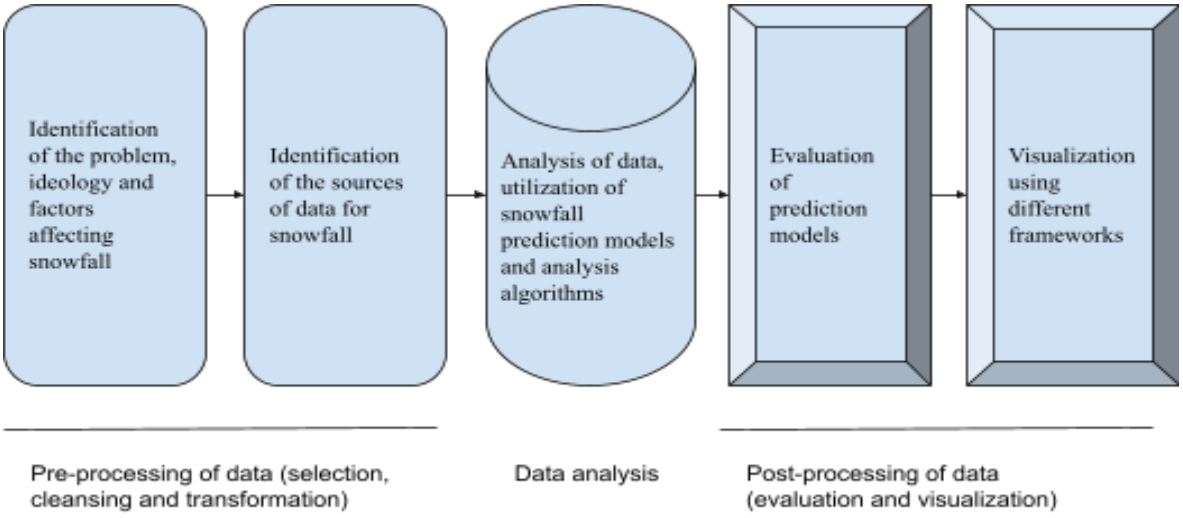

Fig 1: Relationship diagram between the categories for the literature review

S.No	Category	Related work: References	Related work: Description
1		[1]Daniel Eisenberg and Kenneth E. Warner. 2005. Effects of Snowfalls on Motor Vehicle Collisions, Injuries, and Fatalities. American Journal of Public Health 95, 1 (January 2005), 120–124.	The paper estimated the effects of snowfall on US traffic crash rates between 1975 and 2000. The paper linked every fatal crash with the weather report for the day of the occurrence. The paper showed that snow days had fewer fatal crashes than dry days, but more nonfatal-injury crashes and property damage-only crashes. The reason stated was that drivers tended to take some caution due to snow to prevent fatal crashes, but not enough to completely prevent some type of injury. The results show that the first snow day showed a higher number of accidents than subsequent days, particularly from elderly drivers. We took this conclusion as the basis for our decision to consider monthly data.
2		[2]Lennart Quante, Sven N. Willner,	The paper predicts the rate of intensification of snowfall over the northern hemisphere under

	Ideology and factors influencing snowfall and its patterns	Robin Middelanis, and Anders Levermann. 2021. Regions of intensification of extreme snowfall under future warming - Scientific Reports. Nature.	future warming conditions. The paper concludes that extreme snowfall events will become common due to global warming. The paper finds a correlation between global surface temperature and a decrease in snowfall cover in spring. We took this information when considering which parameters to use for our analysis.
3		[3]Hadi Mohammadzadeh Khani, Christophe Kinnard, and Esther Lévesque. 2022. Historical Trends and Projections of Snow Cover over the High Arctic: A Review. MDPI.	Keywords - snow cover duration (SCD), snow cover extent (SCE), snow depth (SD), and snow water equivalent (SWE) The paper studies the change in historical snow patterns in the High Arctic region by observing metrics like SCD, SCE, SD, and SWE. The paper revealed that these metrics changed differently to differences in historical and future changes in precipitation and air temperature. The paper explains in depth the different metrics used for predictions and this helped us in searching for the type of dataset we would need for our analysis.

4		[4]2020. Hotspots of snow cover changes in global mountain regions over 2000–2018. Hotspots of snow cover changes in global mountain regions over 2000–2018 - ScienceDirect.	Keywords - snow water equivalent (SWE), Snow Cover Area (SCA), Snow Cover Duration (SCD), First Snow Day (FSD), Last Snow Day (LSD), Snow Line Altitude (SLA) The study focuses on the decrease in snowfall in mountain regions. The change in temperature at these elevations is twice that observed globally. Hotspots of negative and positive change for parameters like SCA, SCD, FSD, LSD, and SLA are visualized. This paper gave us insight into how elevation is a key factor in how snowfall patterns change and how the latitude of the station-making readings can affect our prediction models.
5	**Snowfall data set**	[5]Jamie L. Dyer and Thomas L. Mote. 2006. Spatial variability and trends in observed snow depth over North America. Geophysical Research Letters 33, 16 (2006).	This paper uses the gridded dataset of daily U.S. and Canadian surface observations from 1960–200. The main goal of this paper is to analyze the spatial and temporal trends of snow depth in the given dataset Daily surface observations have been used which are included in the snow depth data. Linear regression has been used to identify regional trends in snow.

6		[6]Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters 51, 1 (January 2008), 107–113.	This paper discusses how MapReduce helps in cleaning the larger clusters. MapReduce is a programming model that enables the scalability of thousands of servers in a Hadoop cluster. In MapReduce, the user defines a map function that uses a set of data with key/value pairs to generate a new set of data, with elements broken down in a form of a tuple that contains intermediate key/value pairs of that element, and reduce function merges all the intermediate values with the same key. Programs written in this are executed in parallel. The system decides the partitioning of input data, and execution across a set of machines and provides support in machine failure and inter-machine communication.
7		[7]Weather data analysis using spark — An in-memory computing framework. Weather data analysis using spark — An in-memory computing framework | IEEE Conference Publication | IEEE Xplore.	Weather prediction is a complex phenomenon. There are different weather models available to predict the weather accurately. The biggest challenge in predicting weather is to analyze this huge amount of data that is being collected from different weather stations daily. Previously, Hadoop was used in order to analyze this data. In this paper, the author is proposing the Spark implementation for analyzing weather data. Spark is capable of performing in-memory computing which is much more effective than the Hadoop MapReduce.

8		[8]Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. 2019. A Review on Data Cleansing Methods for Big Data. Procedia Computer Science 161, (2019), 731–738.	The paper discusses how a big amount of data can be helpful for organizations in order to predict future decisions based on previous data. The data collected from the different resources are mostly unclean and this data can affect the accuracy of the data, thus we need to clean this data before working on it. Data cleansing is a process of removing unwanted, corrupted, incorrect, or duplicate data to offer better quality data that will be useful to predict the future outcome with much accuracy. Data cleansing is a 5-step process- ● Data analysis ● Definition of transformation workflow and mapping rule ● Verification ● Transformation ● backflow Data cleansing methods included in this paper are - Cleanix, SCARE, KATARA, BigDansing
9		[9]Moon-Soo Song, Hong-Sik Yun, Jae-Joon Lee, and Sang-Guk Yum. 2022. NHESSD - A Comparative Analysis of Machine Learning Algorithms for Snowfall Prediction Models in South Korea. *NHESSD - A Comparative Analysis of Machine Learning Algorithms for Snowfall Prediction Models in South Korea*. Retrieved from https://nhess.copernicus.org/preprints/nhess-2022-118/	In this paper, the performance comparison of 4 machine learning methodologies(MLR, SVR, RFR, XGB) was done. Regression models assisted in predicting a significant amount of snowfall using advanced machine learning algorithms. The usage of grid search and five-fold cross-validation methods helped in enhancing and refining the performance of learning. There was a performance assessment done whose basis was observed and anticipated data. Using various statistical criteria, the RFR model predicted the occurrence of snowfall with the most precision as compared to the other models. This positive outcome points out the potential application of the RFR model to predict extreme snowfall conditions. This can be relevant in our work since we can take inspiration from this model and use the enhancements in this paper to optimize the forecasting technique for our project model.
10		[10]Fraser King, George Duffy, and Christopher G. Fletcher. 2022. A Centimeter-Wavelength Snowfall Retrieval Algorithm Using Machine Learning. *AMETSOC*. Retrieved from https://journals.ametsoc.org/view/jour	In this paper, the utilization of VertiX and ERA5 atmospheric temperature estimates have been highlighted to develop a surface snow accumulation regression model. It is emphasized here that global trends related to snow accumulation can be better

		nals/apme/61/8/JAMC-D-22-0036.1.xml	comprehended using remote sensing snowfall retrievals. The RF model predicts surface snow accumulation with very high accuracy. It uses various event-based training and testing sets for this purpose. It showed stronger performance than the typical snowfall retrieval models that used radar techniques. This work provides evidence that difficulties caused by typical sparse in situ observational networks can be solved easily by nonlinear machine learning-based retrievals. This can help in providing novel and meaningful insights into the global patterns of snow accumulation. We emphasize this work because accumulated snow trends and different phenomena like ecosystem evolution, water resource management techniques, and regional flooding are interconnected strongly. Since our paper also deals with better preparedness in terms of the hazards posed by snowfall, we can take cues from the regression model used in this paper to boost the efficiency of our project model.
11		[11]Paul J. Roebber, Melissa R. Butt, Sarah J. Reinke, and Thomas J. Grafenauer. 2007. Real-Time Forecasting of Snowfall Using a Neural Network. *Weather and Forecasting* 22, 3 (June 2007), 676–684. DOI:https://doi.org/10.1175/waf1000.1	In this paper, 53 snowfall reports have been highlighted and discussed that were gathered in real-time. The main articles of examination for the purpose of snowfall forecasting were the neural network, the surface-temperature-based 676-USDT table, and the climatological snow ratio. It introduces a novel concept of forecast credibility and locates the forecasts within the framework of municipal snow removal. The results highlight that neural networks are best suited for individual events. The forecasting gets higher precision and limited error from the network strategy which effectively compensate for inaccuracies in QPFs. It emphasizes the need for the municipality to be aware of the resources needed for travel readiness on roads. It mentions that an overprediction or an under forecast are both awful since it leads to resource mismanagement. This might result in increased road accidents, worsening travel conditions, and a failure of the core objective. This can spoil overall forecast credibility. We can integrate the idea of using forecast credibility in our project work. This can help in making our prediction model more impactful in terms of its overall

	Snowfall prediction models and algorithms		trustworthiness.
12		[12]IJECRT JOURNAL. 2016. SNOWFALL PREDICTION TECHNIQUES -A STATE OF THE ART REVIEW. *(PDF) SNOWFALL PREDICTION TECHNIQUES -A STATE OF THE ART REVIEW	IJECRT JOURNAL - Academia.edu*. Retrieved from https://www.academia.edu/36738219/SNOWFALL_PREDICTION_TECHNIQUES_A_STATE_OF_THE_ART_REVIEW	This paper reviews the earlier nowcasting methodologies that have been used to predict snowfall. It discusses the models like Intelligent Visibility Meter, machine learning models, and various tools using neural networks. It reviews multiple advanced models and efficient tools and then groups them based on their accuracy and whether they follow a non-decision or decision tree approach. The snow hydrological cycle is a very complex phenomenon to identify and model. We can use the works provided in this paper to learn from the approaches mentioned and then add the optimizations. We can get an estimation of the precision using the approaches listed in this paper that can be used for effective comparative model analysis for our project. This can help us arrive at the most optimal solution.
13		[13]Paolo Sanò, Daniele Casella, Andrea Camplani, Leo Pio D'Adderio, and Giulia Panegrossi. 2022. A Machine Learning Snowfall Retrieval Algorithm for ATMS. *MDPI*. Retrieved from https://www.mdpi.com/2072-4292/14/6/1467	In this paper, the strong points of SLALOM-CT algorithm(a novel ML-based algorithm) have been highlighted. Its implementation majorly consisted of concurrent ATMS and CPR observations for training. It aids in SD, SRE, and SPE. Decision trees and NNs were the strategies that were assessed to find the best performance. The main emphasis was placed on image-based and pixel-based networks and their complexity and depth levels. It was evident that the final SLALOM-CT algorithm demonstrated superior performance in the detection and estimation of snowfall rates (GPROF–ATMS), even when compared to advanced and modern satellite products. The confirmation of its potential can be obtained through radar networks. We can effectively use the observations and approaches used in this paper for fine-tuning our projected model so that the detection of snowfall rate is very precise and builds on historical and current methodologies.
14		[14]Josh Barnwell. 2011. Verification of the Cobb Snowfall Forecasting Algorithm. *"Verification of the Cobb Snowfall Forecasting Algorithm" by Josh Barnwell*. Retrieved from	In this paper, the major emphasis is placed on the Cobb method. This predicts snowfall by utilizing the model data and understanding the relevant snowfall forecasting variables. This generates snowfall amounts for storms in a

		https://digitalcommons.unl.edu/geosciddiss/14	timely manner. It points out that the model errors can be removed by using observational data and a historical model. The high accuracy of the Cobb method can be better realized when it gets compared to the observations. It is proven here that the Cobb Method has a strong precision percentage of 77.7% in observations. It is assumed here to exclude the flaws in the observational data. This paper employs techniques to remove forecast errors which can be used by our project too to make it more resilient. The paper demonstrates that the Cobb method is a positive step and with more research on the factors and improved precision in the model data, the process of forecasting can be made easier. Portions of this can be integrated into our project too and can provide huge advantages in terms of precision and performance. Instead of only focusing on the whole event by utilizing a cumulative snowfall total and snow ratio, the Cobb method can be utilized in our projected model to predict a load of snowfall during specific times of the event. This paper's contribution to providing data for storm events can be really helpful in our project work since storm events pose a huge health hazard and need better preparedness.
15		[15]2017. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives - ScienceDirect*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0957417417303457?fr=RR-2&ref=pdf_download&rr=768ae2de9c6e3149	In this paper, the authors are looking to promote the use of machine learning in the prediction of rainfall. They are using six different algorithms which are: Genetic Programming, Support Vector Regression, Radial Basis Neural Networks, M5 Rules, M5 Model trees, and k-Nearest Neighbors. They are running tests using rainfall time series from 42 different cities having diverse climatic characteristics. What we like about this is that they are evaluating the performance of the ML models against the existing state-of-the-art Markov chain extended to rainfall prediction. By using the latter as the reference, it is assumed to be a meaningful one since it's tried, tested, and widely used.

16	**Evaluation of ML models**	[16]Juan Antonio Bellido-Jiménez, Javier Estévez Gualda, and Amanda Penélope García-Marín. 2021. Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. MDPI.	The aim of the authors of this paper is to fill in missing values in the data collected by the deployment of sensors. The issue addressed is that of predicting those values which might have been missed due to the incapacity of the sensors. They are using models such as SVM (Support Vector Machine), SVR (Support Vector Regression), MLP (Multi-layered Perceptron), and Linear regression to predict missing values in the daily rainfall series of a region in southern Spain. The evaluation by authors includes a correlation between the predicted results of the type of area predicted i.e., inland and coastal. By doing such an evaluation, we can see in which type of region our predictions are more accurate and the opposite as well. We could also try to understand why such behavior exists. This is something we could do in our project.
17		[17]Nawaf Abdulla, Mehmet Demirci, and Suat Ozdemir. 2022. Design and evaluation of adaptive deep learning models for weather forecasting. *Engineering Applications of Artificial Intelligence* 116, (November 2022), 105440.	In this paper, the authors have tried to improve the performance of prediction models by employing many Long term short memory (LSTM). Using this they are trying to find the most accurate and robust model for prediction. Here they are comparing many models such as near regression, Generalized Linear Regression (GLR), decision tree, Gradient-Boosted Tree Regression (GBTR), and random forest. Therefore, to assess the error between the actual value and the predicted value, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), the R-squared coefficient, and Mean Absolute Percentage Error (MAPE) is utilized which seems like the best option for our project as well.

18	**Visualization techniques used for our ML models**	[18]Ajibade, Samuel & Adediran, Anthonia. (2016). An Overview of Big Data Visualization Techniques in Data Mining. 4. 105-113.	These papers discuss some important functions for visualization techniques by using some data mining techniques. Data Visualization helps in selecting makers to effortlessly understand complex thoughts and identify brand-new systems or styles. When Visualization will become interactive, then we are capable of pushing the concepts a little further thereby the usage of technological equipment to grasp more details from graphs and charts, consequently making modifications to the facts that are visible and how such statistics are being processed.
19		[19]Makrufa Hajirahimova and Marziya Ismayilova. 2018. BIG DATA VISUALIZATION: EXISTING APPROACHES AND PROBLEMS. Problems of Information Technology 09, 1 (January 2018), 65–74.	In this paper, the author talks about the existing approaches which are used for visualization techniques and how one can modify these approaches into getting some enhanced output. The traditional approaches used were ineffective and were having some difficulty when it came to dealing with Volume, Velocity, and Variety. In order to come back from this we are gonna use some modern approaches and some examples of the same are as follows: Some of these approaches are Microsoft Power BI, Dashboard, clustergram, Tableau, Plotty, Gephi, and Data-Driven Documents.

20		[20]Samuel Soma Ajibade. An Overview of Big Data Visualization Techniques in Data Mining. (PDF) An Overview of Big Data Visualization Techniques in Data Mining | Samuel Soma Ajibade - Academia.edu	The essential attention of this paper is to provide a quick survey of some of the multi-dimensional visualization strategies which are utilized in information mining, understanding fully properly that the strategies are not limited to the ones that have been mentioned in this paper. Lots of research is being achieved each day pertaining to this aspect and it's far so unfortunate that business sectors are still not getting the precise result they want when seeking to visualize their data using any of the record visualization techniques.

4. COMPONENTS

4.1. Data Processing

4.1.1. Description

We have used the NOAA Global Surface dataset, it's an ISH (Integrated Surface Hourly) based data where USAF Climatology Center provides it. This dataset contains data from 9000 stations around the globe. Few main entities available in this dataset are **Mean Dew Point, Mean Sea Level, Mean Station Pressure, Mean Visibility, Mean Wind speed, Minimum Temperature, Precipitation Amount, Snow Depth, Elevation, Begin**. We have used the data from 2001-2019.

We are using attributes such as temperature, wind speed, mean dew point, precipitation and weather station location for our application.

The step involved in merging the data reported from different weather stations to its location specific to latitude and longitude is using inner merge.

- https://www.kaggle.com/datasets/noaa/gsod
- https://www.kaggle.com/code/ircuscus/global-weather-monitoring-dashboard/data

4.1.2. Data Preprocessing and Cleaning

The Dataset is divided into 2 parts, the first part contains data related to the weather data and the second part contains data related to the mapping. In Mapping the respected station numbers are mapped with their location, i.e there Longitude and Latitude.

For cleaning the dataset, we completely remove any junk values which we found are not contributing towards our final output; this was done either by dropping the respective row values or in some cases the entire column. We were dealt with this, we worked on finding the correlation between these entities inside our dataset. Upon which we found that few things which had a direct correlation were: Dew Point, Wind Speed, Max and Min Temperature, Temperature. Once we have found the correlation, we have combined these two different parts into one.

4.1.3. Integration

We are using libraries in python such as NumPy and Pandas that will be used for data processing. Using these libraries, we are trying to do basic preprocessing by replacing junk values with NaN values.

Provided that the data is structured, we are using Spark SQL functions like merge and filter to merge two dataframes, one which contains weather data along with station number and the other dataframe which contains station number along with its location.

4.1.4. Storage

We are reading our processed csv file as a spark dataframe.

4.2. Prediction and Analysis

Once Data Preprocessing and cleaning is done, we have split the data into 70/30 for training and testing purposes.
We have used the following algorithms for data Modeling purposes. While generating the output of our model or for visualization purposes, we have only used the data for the last 2 years because the ARCGIS Api wouldn't work with 20 years of data. Even while running the model, we tried to run it for the entire dataset but the learning part itself took roughly 10 hours.

```
Correlation to PRCP for  _c0 -0.008497293783364113
Correlation to PRCP for  USAF -0.08285919364652453
Correlation to PRCP for  WBAN -0.04759627872919397
Correlation to PRCP for  YEAR 0.01739869803993931
Correlation to PRCP for  MONTH -0.018013603855147796
Correlation to PRCP for  TEMP -0.11745222440924329
Correlation to PRCP for  DEWP -0.036184829815221
Correlation to PRCP for  WDSP -0.03022177414327929
Correlation to PRCP for  MAX -0.029099478664993492
Correlation to PRCP for  MIN -0.008452812242141125
Correlation to PRCP for  PRCP 1.0
Correlation to PRCP for  DAY 0.0057988125512198835
Correlation to PRCP for  LAT 0.11360050946035001
Correlation to PRCP for  LON -0.04750676947730205
Correlation to PRCP for  ELEV(M) -0.05371110885978155
+--------------------+----+------+--------+----+-----+----+----+--------------------+
|            features|PRCP|   LAT|     LON|YEAR|MONTH| DAY|TEMP|                 LBL|
+--------------------+----+------+--------+----+-----+----+----+--------------------+
|[34.5,23.3,3.8,46...| 0.0|40.747|-122.922|2018|    2|24.5|34.5|LONNIE POOL FIELD...|
|[42.4,32.1,1.6,57...| 0.0|40.747|-122.922|2018|    3|16.0|42.4|LONNIE POOL FIELD...|
|[49.3,35.0,2.4,67...| 0.0|40.747|-122.922|2018|    4|15.5|49.3|LONNIE POOL FIELD...|
+--------------------+----+------+--------+----+-----+----+----+--------------------+
only showing top 3 rows
```

Fig 2: Correlation between different fields on running the ML algorithm

Random Forest Regression:
It's a Supervised Machine Learning Algorithm which is widely used in classification and regression problems. The decision trees are generated using different samples and at the end a majority vote is taken which helps in decision making. One of the reasons why we chose to go ahead with Random forest is because there are a lot of attributes which might affect our decision. So in order to increase the accuracy we decided to use this algorithm. We have used our random state of 10. One thing that we noticed is that the accuracy of Random Forest is lower when it's compared with Gradient Boosted Tree, However, data characteristics might affect the accuracy/Performance. However, while working with Random Forest Model we found that the accuracy of this model was very less compared to Gradient boost. Plus the MSE (Mean Square Error) value was high as well.

Gradient Boost:
XGBoost stands for Extreme Gradient Boosting. This algorithm is mainly used to reduce the bias errors when it's compared with other models. This type of algorithm can be used for both target variables as well as categorical variables. When it's used as a regressor, the cost function is represented by Mean Square Error (MSE) and when it is used as a classifier it's Log Loss.

Mean Squared Error:
The Mean Square Error(MSE) or Mean Squared Deviation of an estimator measures the common of all the squares of errors-that is, the average squared distinction between the envisioned values and the real one's. MSE may additionally talk to the empirical danger, as an estimate of the authentic MSE.

We are using spark.ml to run our machine learning models and evaluators. Both our ML regression models are part of this library. We are using function under the stat class to check correlations between each of our features and precipitation.

```
|[49.3,35.0,2.4,67...| 0.0|40.747|-122.922|2018|    4|15.5|49.  Root M
+--------------------+----+------+--------+----+-----+----+---
only showing top 3 rows

+------------------+----+--------------------+
|        prediction|PRCP|            features|
+------------------+----+--------------------+
| 1.780799961806607| 0.0|[-90.55,9999.9,0....|
|3.2939589759643093| 0.0|[-88.6,9999.9,2.0...|
|0.7185048471987546| 0.0|[-85.2,9999.9,999...|
|0.7185048471987546| 0.0|[-82.2,9999.9,999...|
|0.7185048471987546| 0.0|[-81.050000000000...|
+------------------+----+--------------------+
only showing top 5 rows

Root Mean Squared Error (RMSE) on test data = 21.884
```

Fig 3: Output showing RMSE value on running the ML algorithm on the vast data

4.3. Web Application

In this part, we provide a brief description of the application with respect to the user interface, backend, and frontend implementation.

4.3.1. Back-End

Once the csv file is generated, we are using ArcGIS Python API. We upload the file to ArcGIS online and create a feature layer. We apply this feature layer to create a map on ArcGIS. Once the map is completed, we use ArcGIS JavaScript API to display the map on our custom made webpage.

4.3.2. Front-End

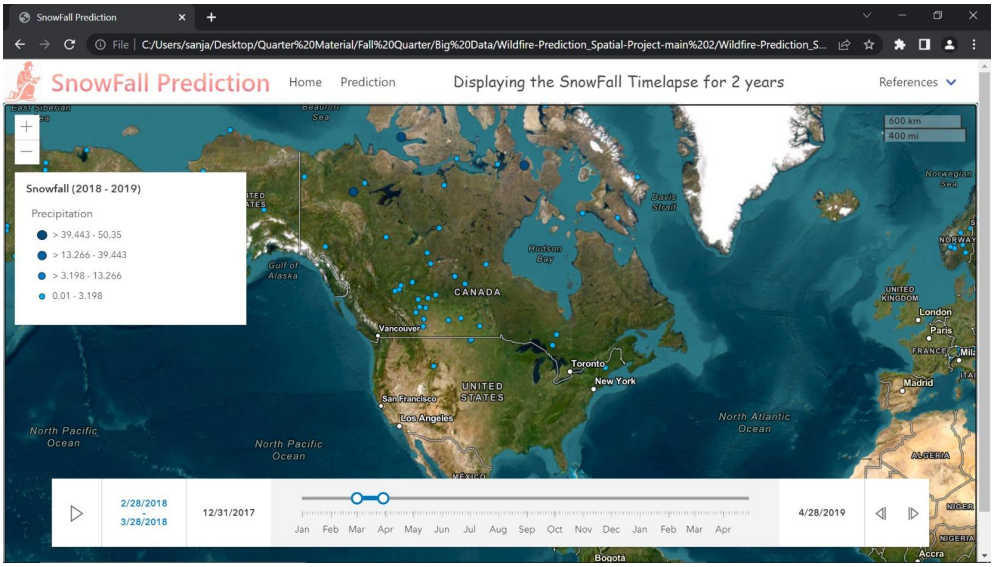

Fig 4: A snapshot of the custom snowfall prediction website integrating timelapse

This front-end UI displays distribution of snowfall for different timeframes. We have developed a custom UI for facilitating this.

4.3.3. Visualization

We have utilized ArcGIS tool to generate the output map which shows the snowfall intensity over a period of time. We have provided an interactive time lapse for the snowfall prediction representation.

We have used 4 different colors to display the snowfall intensity. The legend is provided in the attached image above. Precipitation here represents either snowfall or rainfall and if the temperature is below 22 Fahrenheit then it represents snowfall.

5. EVALUATION

We have used the NOAA Global Surface dataset, it's an ISH (Integrated Surface Hourly) based data where USAF Climatology Center provides it. This dataset contains data from 9000 stations around the globe. We have used the data from 2001-2019. We are using attributes such as temperature, wind speed, mean dew point, precipitation and weather station location for our application.

https://www.kaggle.com/datasets/noaa/gsod

https://www.kaggle.com/code/ircuscus/global-weather-monitoring-dashboard/data

We are using Extreme Gradient Boosting to predict precipitation based on values such as temperature, wind speed, mean dew point, precipitation and weather station location for our application. We have data from 2001-2019 and we are trying to train our model with a random split of 70/30. By the variation of maxIter variable value in the regressor, we are comparing it to the RMSE values to check which is the ideal maxIter value suitable for our prediction. It seems like the max iteration value affects the RMSE inversely but the effect reduces with the increase and seems like it will flatten at a point. The reason we did this analysis is to see if the time trade-off by increasing maxIter to the improvement in RMSE value. But the time increase is not viable if we are looking to scale and train with even more years of data. Our RMSE value is also not ideal to trust the predictions. Even though research has shown that Gradient Boost is good for prediction, it shows that there might be more improvement that can be made by using more fields and attributes.

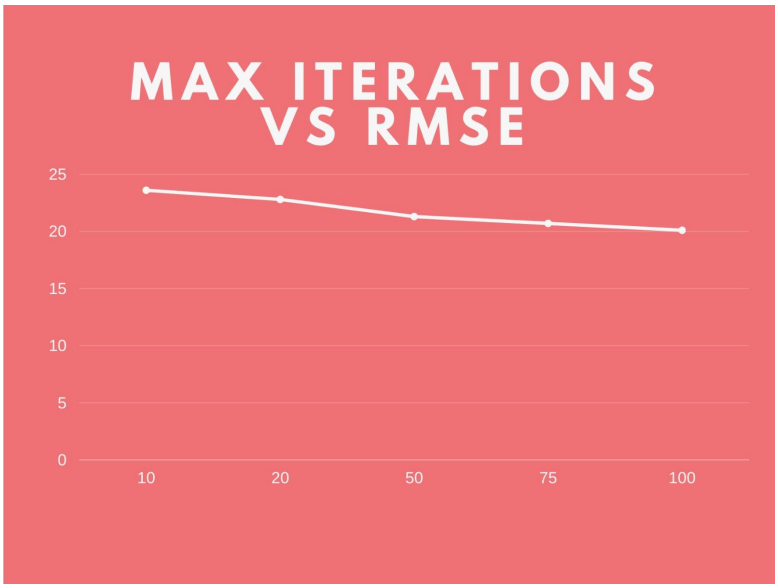

Fig 5: Evaluation relation obtained for different maxIterations and RMSE

6. CONCLUSION

The results obtained from this project were able to provide a comparative analysis of the historical patterns of snowfall data using big data technologies and we are visualizing it using ArcGIS. We take the data from multiple years, extract the tar file for the dataset, perform random splitting, evaluate gradient boost regression test data and then find the Root Mean Squared Error (RMSE) on test data. Our machine learning model was able to predict with a Root Mean Squared Error (RMSE) on test data equal to 21.7968. With better models, we can understand the changes in snowfall patterns due to the temperature even more precisely. As per this project, we visualize the data using ArcGIS where we show the snowfall prediction for multiple years. We have taken inspiration from RMSE and Gradient-Boosted Tree Regression from the earlier works, especially [17]. We effectively utilized big data technologies to process huge datasets and built a prediction model. This project can be further expanded to the global datasets where processing and prediction can be done on a larger scale involving multiple countries so that it can result in a world wide snowfall prediction. Our hope is that this can be utilized by governments bodies to better prepare themselves for weather disasters.

7. AUTHOR CONTRIBUTIONS

Name	**Major Contributions**
Puneet Singhania	● Implementation of Spark ● Front-End development for the visualization part ● Evaluation of the model ● Literature survey of the category of Snowfall prediction models and algorithms
Rigved Patil	● Implemented the Machine Learning Algorithms ● Front-end development for the visualization part

	• Literature survey of the category of Visualization techniques used for our ML models
Ritesh Singh	• Implementation of Spark • ArcGis Implementation • Literature survey of the category of Snowfall data set.
Sanjana Senthilkumar	• ArcGIS Implementation • Frontend and Backend API connection with ArcGIS • Dataset identification • Implementation of Spark • Literature Survey for Ideology and factors influencing snowfall and its patterns
Shadrush Swaroop	• Implemented the Machine Learning Algorithms • Front-End Development for the visualization part • Dataset cleaning • Literature survey for Evaluation of ML models

8. REFERENCES

[1] Climate Change Indicators: Snow Cover. Retrieved October 12, 2022 from https://www.epa.gov/climate-indicators/climate-change-indicators-snowfall

[2] National Centers for Environmental Information. National Centers for Environmental Information (NCEI). (n.d.). Retrieved October 13, 2022, from https://www.ncei.noaa.gov/

[3] Public Safety Canada. 2022. Winter storms. (August 2022). Retrieved October 15, 2022 from https://www.publicsafety.gc.ca/cnt/mrgnc-mngmnt/ntrl-hzrds/wntr-strm-en.aspx

[4] "United States Department of Labor." Winter Weather - Hazards/Precautions| Occupational Safety and Health Administration, https://www.osha.gov/winter-weather/hazards

[5] "Data Access." National Centers for Environmental Information (NCEI), https://www.ncei.noaa.gov/access/search/dataset-search?observationTypes=Land+Surface&startDate=2021-01-06T00%3A00%3A00&endDate=2021-01-06T23%3A59%3A59

[6] Daniel Eisenberg and Kenneth E. Warner. 2005. Effects of Snowfalls on Motor Vehicle Collisions, Injuries, and Fatalities. American Journal of Public Health 95, 1 (January 2005), 120–124.

[7] Lennart Quante, Sven N. Willner, Robin Middelanis, and Anders Levermann. 2021. Regions of intensification of extreme snowfall under future warming - Scientific Reports. Nature.

[8] Hadi Mohammadzadeh Khani, Christophe Kinnard, and Esther Lévesque. 2022. Historical Trends and Projections of Snow Cover over the High Arctic: A Review. MDPI.

[9] 2020. Hotspots of snow cover changes in global mountain regions over 2000–2018. Hotspots of snow cover changes in

global mountain regions over 2000–2018 - ScienceDirect.

[10] Jamie L. Dyer and Thomas L. Mote. 2006. Spatial variability and trends in observed snow depth over North America. Geophysical Research Letters 33, 16 (2006).

[11] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters 51, 1 (January 2008), 107–113.

[12] Weather data analysis using spark — An in-memory computing framework. Weather data analysis using spark — An in-memory computing framework | IEEE Conference Publication | IEEE Xplore.

[13] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. 2019. A Review on Data Cleansing Methods for Big Data. Procedia Computer Science 161, (2019), 731–738.

[14] Moon-Soo Song, Hong-Sik Yun, Jae-Joon Lee, and Sang-Guk Yum. 2022. NHESSD - A Comparative Analysis of Machine Learning Algorithms for Snowfall Prediction Models in South Korea. NHESSD - A Retrieved from https://nhess.copernicus.org/preprints/nhess-2022-118/

[15] Fraser King, George Duffy, and Christopher G. Fletcher. 2022. A Centimeter-Wavelength Snowfall Retrieval Algorithm Using Machine Learning. AMETSOC. Retrieved from https://journals.ametsoc.org/view/journals/apme/61/8/JAMC-D-22-0036.1.xml

[16] Paul J. Roebber, Melissa R. Butt, Sarah J. Reinke, and Thomas J. Grafenauer. 2007. Real-Time Forecasting of Snowfall Using a Neural Network. Weather and Forecasting 22, 3 (June 2007), 676–684. DOI:https://doi.org/10.1175/waf1000.1

[17] IJECRT JOURNAL. 2016. SNOWFALL PREDICTION TECHNIQUES -A STATE OF THE ART REVIEW. (PDF) SNOWFALL PREDICTION TECHNIQUES -A STATE OF THE ART REVIEW | IJECRT JOURNAL - Academia.edu. Retrieved from https://www.academia.edu/36738219/SNOWFALL_PREDICTION_TECHNIQUES_A_STATE_OF_THE_ART_REVIEW

[18] Paolo Sanò, Daniele Casella, Andrea Camplani, Leo Pio D'Adderio, and Giulia Panegrossi. 2022. A Machine Learning Snowfall Retrieval Algorithm for ATMS. MDPI. Retrieved from https://www.mdpi.com/2072-4292/14/6/1467

[19] Josh Barnwell. 2011. Verification of the Cobb Snowfall Forecasting Algorithm. "Verification of the Cobb Snowfall Forecasting Algorithm" by Josh Barnwell. Retrieved from https://digitalcommons.unl.edu/geoscidiss/14

[20] 2017. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives - ScienceDirect. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0957417417303457?fr=RR-2&ref=pdf_download&rr=768ae2de9c6e3149

[21] Juan Antonio Bellido-Jiménez, Javier Estévez Gualda, and Amanda Penélope García-Marín. 2021. Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. MDPI.

[22] Nawaf Abdulla, Mehmet Demirci, and Suat Ozdemir. 2022. Design and evaluation of adaptive deep learning models for weather forecasting. Engineering Applications of Artificial Intelligence 116, (November 2022), 105440.

[23] Ajibade, Samuel & Adediran, Anthonia. (2016). An Overview of Big Data Visualization Techniques in Data Mining. 4.

105-113.

[24] Makrufa Hajirahimova and Marziya Ismayilova. 2018. BIG DATA VISUALIZATION: EXISTING APPROACHES AND PROBLEMS. Problems of Information Technology 09, 1 (January 2018), 65–74.

[25] Samuel Soma Ajibade. An Overview of Big Data Visualization Techniques in Data Mining. (PDF) An Overview of Big Data Visualization Techniques in Data Mining | Samuel Soma Ajibade - Academia.edu

[26] Roebber, P. J., Bruening S. L. , Schultz D. M. , and Cortinas J. V. Jr., 2003: Improving snowfall forecasting by diagnosing snow density. Wea. Forecasting, 18 , 264–287