American International University-Bangladesh (AIUB)

Long term Climate Data Analysis Using Machine Learning Algorithms

Sajjad Hossein (20-43639-2)
Fardin Ahsan (17-33941-1)
Tahajjod Bhuiya (18-38566-2)
Noman Bhuyan (20-43830-2)

A Thesis submitted for the degree of Bachelor of Science (BSc)
in Computer Science and Engineering (CSE) at
American International University Bangladesh in July, 2021

Faculty of Science and Technology (FST)

Abstract

This thesis explores long-term climate data analysis using machine learning algorithms to uncover significant trends and inter-variable relationships in climate behavior over time. Focusing on monthly rainfall data, this research employs five machine learning algorithms: K-Nearest Neighbors (KNN) for clustering similar patterns within climate data, Bayesian Belief Networks (BBN) to identify dependencies among climate variables, Trend Analysis to observe seasonal shifts and overall trends, Autoregressive Integrated Moving Average (ARIMA) for analyzing temporal correlations, and Seasonal Decomposition of Time Series (STL Decomposition) to extract seasonal and long-term components of the data. These techniques were picked because they offer a thorough grasp of slow climate change without putting a premium on forecast accuracy. The data preprocessing steps included managing missing values and handling duplicate entries to ensure data consistency. KNN clustering was used to group similar patterns across the dataset, revealing reoccurring climate behaviors and patterns. BBN analysis highlighted important dependencies between rainfall and other climate variables, which could act as potential predictors of regional climate variability. Trend Analysis and STL Decomposition enabled the identification of seasonal and cyclical shifts, showing periodic and gradual changes within the climate data. Finally, ARIMA modeling provided insights into temporal trends, emphasizing time-based dependencies that further confirm long-term directional shifts in rainfall patterns. The results show recognizable changes in rainfall trends and highlight a better comprehension of inter-variable relationships that can guide adaptation plans and regional climate assessments. By highlighting the importance of trend and dependency analysis over exact predictive accuracy, this study advances the field of climate science and provides insightful information for scholars and decision-makers involved in environmental planning and climate adaptation. The findings lay the groundwork for further studies in climate trend analysis and highlight the value of a multifaceted approach in comprehending intricate climate dynamics.

Declaration by author

(All candidates to reproduce this section in their thesis verbatim)

This thesis is composed of our original work, and contains no material previously published or written by another person except where due reference has been made in the text. We have clearly stated the contribution of others to our thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in our thesis. The content of our thesisis the result of work we have carried out since the commencement of Thesis / Software project.

We acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate we have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Name of the Author 1
Student ID
Computer Science & Engineering
Faculty of Science & Technology

Name of the Author 2
Student ID
Computer Science & Engineering
Faculty of Science & Technology

Name of the Author 3
Student ID
Computer Science & Engineering
Faculty of Science & Technology

Name of the Author 4
Student ID
Computer Science & Engineering
Faculty of Science & Technology

Approval

(All candidates must edit this page)

The thesis titled **“Thesis title goes here”** has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirements for the degree of Master of Science in Computer Science on **(date of defence)** and has been accepted as satisfactory.

Supervisor’s Name
Rank & Position

Name of the External
Rank & Position

Name of the Head (UG/PG)
Rank & Position

Prof. Dr. Dip Nandi
Associate Dean
Faculty of Science and Technology
American International University-Bangladesh

Mashiour Rahman
Sr. Associate Professor & Dean-in-charge
Faculty of Science and Technology
American International University-Bangladesh

Publications included in this thesis

If you choose to include publications as part of your thesis use this section to detail accepted or in press publication(s) using the standard citation format for your discipline.

Papers submitted for publication and awaiting review should appear in the next section, **Submitted manuscripts included in this thesis**.

If you have not included any of your publications in the thesis then state "No publications included".

Example:

1. [Dum, 2021] **Your Name**, Co-author 1, and Final Author, Title of your paper, *Journal*, Issue, Number, Year
2. [Dum, 2021] **Your Name**, Co-author 1, and Final Author, Title of your paper, *Journal*, Issue, Number, Year

Submitted manuscripts included in this thesis

List manuscript/s submitted for publication here. As described above for **Publications included in the thesis**, on the page immediately preceding the chapter that includes the submitted manuscript, in no more than one (1) page, detail your contribution to the authorship if you are not the sole author.

If you have no submitted manuscripts from your candidature then state "No manuscripts submitted for publication".

Example:

1. [Dum, 2021] **Your Name**, Co-author 1, and Final Author, Title of your paper, submitted to *Journal* on 4th June 2018.

Other publications during candidature

List other publications arising during your candidature using the standard citation format for your discipline. Divide your publications into sub-sections as appropriate in your discipline *e.g.* peer-reviewed papers, book chapters, conference abstracts. Papers submitted for publication and awaiting review are not considered publications and cannot be included in this section.

If you have no publications from your candidature then state "No other publications".

Example:

Conference abstracts

1. [Dum, 2021] **Your Name**, Co-author 1, and Final Author, Title of your conference paper, *Proceedings of Conference*, other details.

Book chapters

1. [Dum, 2021] **Your Name**, Co-author 1, and Final Author, Title of your chapter, Book, editor, *etc.*.

Research involving human or animal subjects

All reserch involving human or animal subjects requires prior ethical review and approval by an independent review committee. At UQ, the relevant committee for research involving human subjects is the Human Ethics Unit and the relevant committee for research involving animal subjects is the relevant Animal Ethics Committee. Please provide details of any ethics approvals obtained including the ethics approval number and name of approving committees. A copy of the ethics approval letter must be included in the thesis appendix.

If no human or animal subjects were involved in this research please state: "No animal or human subjects were involved in this research".

Contributions by authors to the thesis

List the significant and substantial inputs made by different authors to this research, work and writing represented and/or reported in the thesis. These could include significant contributions to: the conception and design of the project; non-routine technical work; analysis and interpretation of researchdata; drafting significant parts of the work or critically revising it so as to contribute to the interpretation.

	Sajjad Hossein	**Full Name**	**Full Name**	**Full Name**	**Contribution (%)**
	20-43639-2	*xx-xxxxx-x*	*xx-xxxxx-x*	*xx-xxxxx-x*	
Conceptualization					100 %
Data curation					100 %
Formal analysis					100 %
Investigation					100 %
Methodology					100 %
Implementation					100 %
Validation					100 %
Theoretical derivations					100 %
Preparation of figures					100 %
Writing – original draft					100 %
Writing – review & editing					100 %

If your task breakdown requires further clarification, do so here. Do not exceed a single page.

Acknowledgments

First of all, I would like to thank almighty Allah, for his grace in accomplishing our thesis timely and our families for giving us mental and financial support throughout the whole bachelors. We would like to express our gratitude to the Faculty of Science & Technology to keep thesis credit in the curriculum of the graduation program and giving us a scope of tasting the flavor or research work with our interest.

We also want to express our gratitude to our supervisor, Dr. Dip Nandi sir, who gave us motivation, guidance, and control for this venture. He has decent knowledge of programming advancement, and his personal circumstances, perpetual assistance made the task do able.

Keywords

Climate data, machine learning, trend analysis, rainfall patterns, climate relationships, KNN, Bayesian Belief Network, ARIMA, STL decomposition, long-term climate

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Mention all the abbreviations and the different symbols that is used in this document.

Abbreviations	
BBN	Bayesian Belief Network
CSE	Computer Science and Engineering
KNN	K-Nearest Neighbors
ARIMA	Autoregressive Integrated Moving Average
STL	Seasonal-Trend Decomposition of Time Series
ML	Machine Learning
AI	Artificial Intelligence
CS	Comma-Separated Values
RMSE	Root Mean Square Error
MSE	Mean Square Error
etc.	*etc.*

Symbols	
ρ	Density of a variable
σ	Standard deviation
Σ	Summation
$\sum$	Sum of terms
etc.	*etc.*

Chapter 1

Introduction

1.1 Thesis Topic

Long Term Climate Data Analysis Using Machine Learning Algorithms.

1.2 Introduction

Globally, climate change presents serious problems for ecosystems and societies, affecting biodiversity, public health, water resources, and agriculture. In recent decades, the study of climate data has gained immense significance due to the pressing need to understand climate change and its impact on the environment. Long-term climate data analysis can help identify gradual changes in weather patterns, particularly in rainfall, temperature, and seasonal variations. As climate variables are inherently complex and interdependent, machine learning (ML) algorithms offer powerful tools for identifying trends and analyzing relationships within this data. Understanding and addressing the effects of climate variability now requires the ability to evaluate and interpret long-term climate data. A strong methodology for examining intricate, sizable climate datasets is provided by machine learning, which helps researchers spot trends and connections that conventional approaches might miss. This study intends to provide insights into changing climate trends by using machine learning techniques on historical climate data. It emphasizes the significance of comprehending slow changes in weather patterns rather than concentrating only on prediction accuracy.

1.3 Motivation

The motivation for this research arises from the need for accurate and nuanced analyses of long-term climate trends that can inform sustainable development and environmental policy. While numerous studies have focused on predicting short-term weather conditions or specific climate indicators, fewer investigations have concentrated on exploring long-term trends and relationships within climate variables. Effective climate adaptation and mitigation strategies depend on an understanding of these trends, according to the Intergovernmental Panel on Climate Change (IPCC) and other international environmental organizations. By applying machine learning to the analysis of climate data, this study will advance this field by providing insights into subtle changes that might not be immediately noticeable, like variations in seasonal rainfall patterns.

1.4 Objective

The primary objective of this study is to analyze long-term climate trends and relationships in climate data, focusing on monthly rainfall and other related climate variables. This thesis utilizes five machine learning algorithms: K-Nearest Neighbors (KNN), Bayesian Belief Network (BBN), Trend Analysis, ARIMA, and STL Decomposition. The specific objectives are outlined as follows:

1. Identify Long-Term Climate Trends: Analyze historical climate data to uncover long-term trends, particularly focusing on rainfall and seasonal patterns.
2. Examine Relationships Between Climate Variables: Use machine learning algorithms to study dependencies and relationships among key climate variables, such as rainfall and temperature, to better understand how these factors interact over time.
3. Utilize Machine Learning Algorithms to Extract Insight: Investigate different patterns in the climate data using K-Nearest Neighbors (KNN), Bayesian Belief Network (BBN), Trend Analysis, ARIMA, and STL Decomposition, prioritizing descriptive insights over predictive accuracy.
4. Enable Fundamental Understandings for Research and Policy: Focus on long-term climate shifts and gradual changes to produce findings that can help researchers, policymakers, and environmental planners develop strategies for climate adaptation.
5. Highlight the Value of Non-Predictive Climate Analysis: Provide a distinct viewpoint in climate science research by highlighting the value of examining climate trends and relationships over time as an adjunct to predictive modeling.

1.5 Orientation

A thorough literature review of previous research on climate trend analysis and the application of machine learning to climate research is given in Chapter 2. The data set, preprocessing procedures, and the justification for each algorithm selected for this investigation are described in Chapter 3. With an emphasis on their use in the analysis of climate data, Chapter 4 describes the methodological approach taken for each machine learning algorithm. The results are presented in Chapter 5, which also analyzes long-term trends, clusters, and dependencies between variables. Chapter 6 concludes by discussing the findings' implications and offering suggestions for additional research.

Chapter 2

Literature review

Predictive modeling and machine learning (ML) have emerged as crucial instruments in recent years for examining climate data, forecasting environmental occurrences, and identifying trends in climate behavior. According to studies, machine learning (ML) techniques are especially useful when working with time series data because they can handle intricate relationships between variables, identify non-linear patterns, and provide strong predictive power. The application of different machine learning techniques in climate-related domains is examined in this literature review, with particular attention paid to algorithms that are pertinent to our investigation: Bayesian Belief Network (BBN), K-Nearest Neighbors (KNN), ARIMA, STL Decomposition, and other models. By using these methods for long-term climate trend analysis, it also draws attention to existing gaps that our research seeks to fill.

Research on time series prediction models, especially ARIMA and SARIMA, demonstrates that statistical methods are effective in detecting linear trends within stationary data. For instance, Selmy et al. applied ARIMA to predict temperature data and found that while the model performs well with linear data, it struggles with non-linear and complex datasets [1]. Similarly, Khan et al. compared models like Support Vector Machines (SVM), Artificial Neural Networks (ANN), and KNN for drought prediction in Pakistan. They concluded that while SVM was most accurate in drought classification, KNN was effective in identifying general patterns but less so for precise forecasting [3].

In order to evaluate risk factors and forecast events based on past data, flood prediction models frequently use machine learning techniques. Razali et al. modeled flood risks using BBN, KNN, and SVM in a CRISP-DM framework, highlighting BBN's benefit in managing class imbalances via SMOTE resampling. Their results highlight the usefulness of BBN in dependency modeling, which makes it advantageous for climate variables like temperature and precipitation [2].

Since drought prediction depends on a variety of climate factors, including temperature, humidity, and precipitation, it presents special difficulties. In comparison to KNN and ANN, Khan et al. achieved superior drought classification by utilizing feature selection techniques such as Recursive Feature Elimination (RFE) in SVM. Despite its superiority in pattern recognition over accuracy, this result demonstrates the usefulness of KNN for environmental prediction tasks [3].

Because it makes it easier to spot long-term changes in factors like temperature, precipitation, and streamflow, trend analysis is crucial in climate studies. Trends in climate time series are frequently extracted using methods like seasonal decomposition and moving averages. Planning for resources requires an understanding of how climatic changes appear over long time periods, which trend analysis helps to provide. Hosseinzadeh et al. [4] used trend analysis to predict streamflow and found that temperature and precipitation are important climate variables that impact the dynamics of water flow in river systems.

An effective non-parametric algorithm for classification and regression in datasets with non-linear relationships is kNN. Based on historical data, kNN has demonstrated efficacy in hydrology and environmental science in forecasting variables such as streamflow and water quality. Its simplicity and adaptability, especially in scenarios with few assumptions, make its application to climate data promising. In a study that used kNN to predict water quality, it was found to be useful for controlling seasonal and non-linear variations in river environments [5]. The selection of neighbors and distance metrics, however, can have an impact on kNN's performance, which could have an impact on outcomes in high-dimensional climate datasets.

BBNs are useful for evaluating the uncertainties in climate projections because they are probabilistic graphical models that depict the relationships between climate variables. Because BBNs can simulate intricate relationships within hydrological systems, researchers can better comprehend how various climatic factors affect outcomes like precipitation and streamflow. Significant wave heights were modeled using BBNs in [6], showcasing their ability to manage complex probabilistic relationships in climate data. By combining real-time data with historical climate information, BBNs can enhance the robustness and interpretability of long-term climate projections.

Wang et al. applied STL decomposition in predicting the remaining useful life of aeroengines, which has broader implications in climate data analysis. By isolating seasonal, trend, and residual components, STL decomposition allows for the precise identification of periodic climate variations and long-term trends. This approach enables more granular modeling, which is essential for detecting both recurring seasonal changes and gradual shifts over time [7].

In their comparative study, Marwa Farouk et al. investigated the application of kNN in forecasting weather patterns, specifically for atmospheric phenomena like dust storms. Their study proved that kNN could efficiently handle large amounts of meteorological data by classifying various weather conditions using historical datasets in conjunction with decision trees and Naïve Bayes. This approach is a reliable tool for climate analysis since it has been demonstrated to be successful in revealing hidden patterns in sizable databases [8].

Shah et al. used ARIMA for time series forecasting. Because it relies on historical values, the ARIMA model has shown promise in predicting seasonal climate trends and efficiently captures temporal dependencies in precipitation data. This model has proven to be highly accurate in meteorological applications and is still widely used for seasonal forecasting [9].

A study by M. Farouk et al. [7] presents a data mining approach for atmospheric dust forecasting using Decision Trees, K-Nearest Neighbors (KNN), and Naive Bayes algorithms. The study's methodology was divided into three primary phases: predicting atmospheric dust levels, training the model using historical weather data, and extracting features from meteorological data (such as temperature, wind speed, and humidity). Of the three models that were tested, the Decision Tree algorithm performed better than both KNN and Naive Bayes, classifying atmospheric dust levels with an accuracy of 89.5%. The Decision Tree's resilience in managing high-dimensional, noisy data—which is frequently the case with meteorological datasets—is demonstrated by this outcome. Farouk et al. concluded that Decision Tree algorithms are well-suited for climate data applications where interpretability and accuracy are both critical [7].

In another study, Shah et al. [8] investigated the effectiveness of Random Forest, Decision Trees, and KNN for classifying rainfall levels. Their approach included a preprocessing phase to handle missing data, followed by feature engineering to enhance model performance. The experimental findings showed that Random Forest outperformed Decision Tree and KNN models, achieving the highest classification accuracy of 92.3%. The ensemble nature of Random Forest, which lowers overfitting and enhances model generalizability, is responsible for this better performance. According to Shah et al., the Random Forest model is especially well-suited for rainfall classification tasks in climate studies because of its capacity to handle massive data sets and capture intricate relationships between climate variables [8].

Mohsin and Gough conducted a trend analysis of long-term temperature variations in the Greater Toronto Area (GTA), employing non-parametric statistical methods such as the Mann–Kendall test and Theil-Sen slope estimator. Urbanization and temperature rise over the previous few decades may be related, according to the study, which revealed statistically significant increases in annual and seasonal mean temperatures, especially during the winter. This approach, which included data from a variety of urban, suburban, and rural weather stations, demonstrated how urbanization had a distinct effect on the climate in different GTA locations and showed a significant warming trend that may be attributed to urban growth [9].

Tian et al. investigated the temporal consistency of global NDVI datasets from multiple sensor systems, focusing on minimizing artifacts due to sensor shifts in long-term trend analysis. Using NDVI datasets like GIMMS3g and VIP3, the study employed structural change detection algorithms to find discrepancies associated with sensor transitions, which were especially noticeable in dry and semi-arid locations. To increase data dependability, which is essential for accurately identifying vegetation trends and changes over time, the authors suggest improved preprocessing techniques [10].

Vijay and Varija used methods including the Mann-Kendall test, modified Mann-Kendall tests, and creative trend analysis to analyze 119 years of climate data in order to do a machine learning-based evaluation of long-term climate variability in Kerala, India. According to their research, mean and maximum temperatures at the seasonal and yearly scales have been rising while annual and monsoon rainfall has been declining. The findings, which highlight possible difficulties for regional water management in the context of climate change, show notable diversity in rainfall and temperature patterns throughout Kerala and include clustering and principal component analysis [11].

From 1966 to 2005, Tabari et al. [12] examined long-term patterns in Iran's precipitation and air temperature. In order to identify trends in annual maximum, minimum, and mean temperatures as well as precipitation, the study used the Mann-Kendall, Mann-Whitney, and Mann-Kendall rank statistic tests. Temperature variables showed a notable warming trend, with mean, maximum, and minimum temperatures increasing by +0.412 °C, +0.452 °C, and +0.493 °C per decade, respectively. Precipitation patterns, however, differed throughout the region, exhibiting both increases and decreases. These results highlight the complexity of the effects of climate change in Iran across various climatic variables and geographical areas.

A thorough analysis of time series analysis methods unique to climate studies was given by Mudelsee [13]. In order to handle issues like non-normal distributions, serial dependence, and irregular time intervals in climate data, the book highlights the significance of error estimation using bootstrap techniques. This method improves the accuracy of trend and variability estimates in climate data by using Monte Carlo simulations and bootstrap resampling. By precisely measuring uncertainties in time series estimations, this methodology supports better predictions and provides a strong framework for examining intricate climate patterns.

Sayemuzzaman et al. [14] analyzed temperature trends across North Carolina over 60 years, using the Mann-Kendall test and Theil-Sen slope estimator. The study discovered a warming trend in average temperature in coastal areas and a cooling trend in mountainous regions. The Sequential Mann-Kendall test revealed shifts in minimum and maximum temperatures around 1970 and 1960, respectively. The study identifies significant regional differences in temperature trends, which are linked to geographic features like mountains and coastal proximity. The findings help to better understand localized climate variability, which is important for regional climate adaptation and policy planning.

A study by Wang et al. explores a technique for predicting the Remaining Useful Life (RUL) of aero-engines using a time-series decomposition model paired with similarity measurements. The method uses fuzzy clustering to identify degradation stages and health indicators (HIs) to quantify engine health. This method identifies high-similarity segments in an engine's historical degradation data by decomposing multidimensional time-series data into trend, seasonal, and residual components. When tested on NASA's aero-engine datasets, this method achieved an MSE of 528, demonstrating superior accuracy over other comparative approaches in the literature, particularly for managing safety-critical predictions [15].

Rahman et al. present a machine learning fusion framework for rainfall prediction designed specifically for smart city applications. This method combines four classifiers—Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—with a fuzzy logic-based fusion layer to improve predictive accuracy. Using 12 years of historical weather data, the system performs pre-processing steps such as data cleaning and normalization prior to classification. The results show that the fusion framework outperforms standalone classifiers, making it suitable for real-time applications that require high accuracy, such as weather forecasting in urban areas [16].

Bayma and Pereira investigate machine learning techniques for reconstructing missing climate data in time-series datasets from Minas Gerais, Brazil. This study uses neural networks, bagged regression trees, and random forests to estimate missing values in various climate attributes. Using statistical analysis and visualization techniques, the authors determine that random forest is the most effective approach for demonstrating resilience in scenarios with a high rate of missing data. The framework successfully restores data continuity in historical records, laying a solid foundation for future climate modeling and forecasting [17].

Evans and Shen [18] proposed a method combining remotely sensed phenology and seasonal climate metrics with machine learning to estimate wheat yield. They derived phenological metrics, including the timing and duration of growth stages, by estimating the spatially weighted growth curve (SWGC) from Landsat NDVI data. They compared six models, including deep learning and support vector regression (SVR), and discovered that SVR and deep learning outperformed the others, with an RMSE of 0.32 t/ha and an R2 of 0.68. Precision agriculture decision-making through comprehensive spatial and temporal yield mapping is made possible by this method, which showed that long-term yield hindcasts at 30-meter spatial resolution are achievable [18].

Rasuly et al. [19] analyzed the time series of synoptic pressure systems affecting Iran's seasonal precipitation using Mann-Kendall trend analysis. Significant shifts were found in the Mediterranean and Sudanese low pressure systems, which weakened by 3.9 and 2.4 mb in the winter and spring, respectively, according to the study. Rainfall and cyclogenesis decreased as a result of this weakening. In a similar vein, monsoon low pressures over Pakistan decreased by 3.8 mb, which made the weather in southeast Iran drier. These results demonstrate how synoptic patterns influence long-term climate variability and how they affect regional trends in precipitation [19].

Dimri et al. [20] applied a Seasonal ARIMA (SARIMA) model to analyze and forecast precipitation and temperature trends in the Bhagirathi River Basin in India. Using data from 1901 to 2000, SARIMA(0,1,1)(0,1,1)12 was identified as the optimal model for precipitation, and SARIMA(0,1,0)(0,1,1)12 for temperature. Forecasts for 2001–2020 revealed that while SARIMA was good at capturing seasonal patterns, it had a tendency to overestimate instances of intense rainfall. The findings highlight the applicability of SARIMA for comprehending climate factors in areas with intricate topographies and notable hydrological variability [20].

Chapter 3

Methods

The methodology used to analyze long-term climate data with a particular emphasis on rainfall patterns using machine learning algorithms is explained in the methods chapter. Starting with data collection, preprocessing, feature extraction, and machine learning techniques, this study integrates several steps. Trend forecasting and result analysis come next. In order to accomplish the goals of the research finding patterns, evaluating dependencies, and predicting long-term trends, the methodology guarantees an organized workflow.

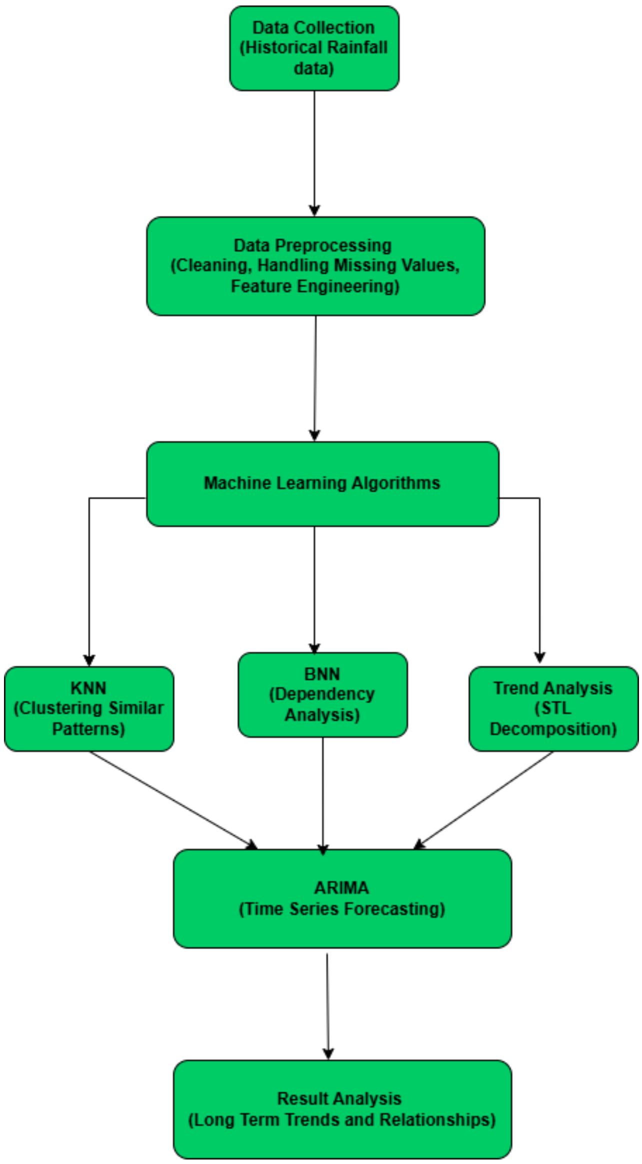

Figure 3.1: Flow chart of proposed methodology

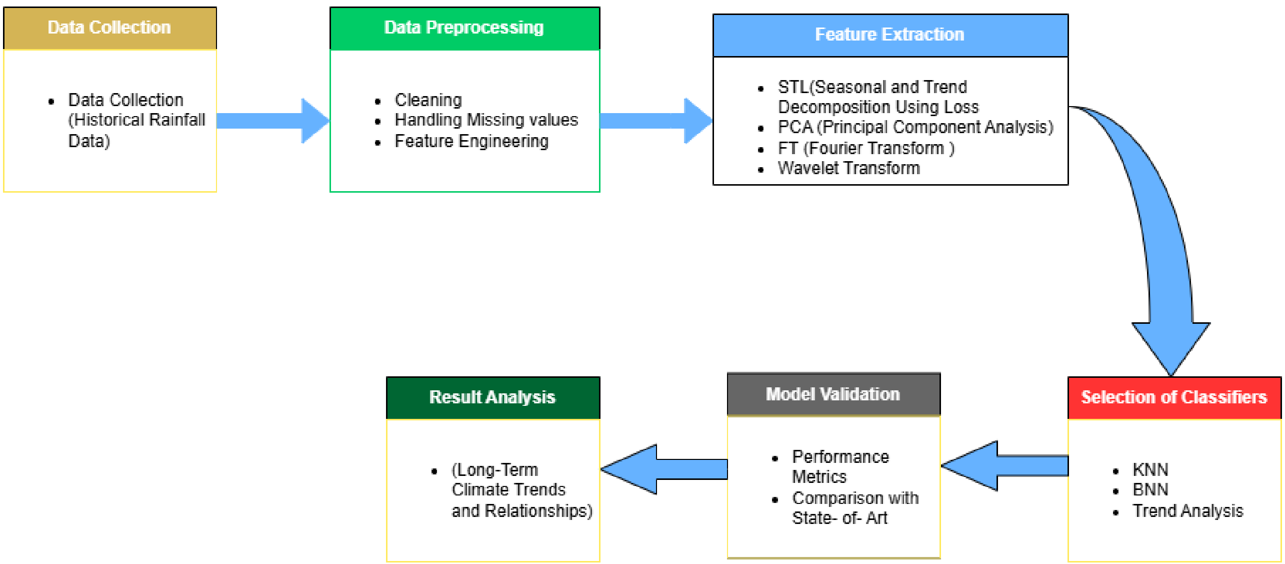

Figure 3.2: Diagram of proposed methodology

3.1 Data Collection

The dataset used for this study comprises historical monthly rainfall data, which was collected from trusted meteorological sources and climate databases over an extended period. The data set captures rainfall measurements recorded at regular monthly intervals across a specific geographic region. Each entry in the dataset includes the total monthly rainfall, along with temporal information (e.g., year and month). This comprehensive dataset provides a detailed temporal overview of rainfall variations over the years, making it suitable for identifying seasonal trends, anomalies, and long-term climate shifts. The data serves as the foundation for analyzing patterns and relationships in climate behavior, ensuring the study's relevance in understanding rainfall variability. By leveraging real-world historical data, this research aims to uncover significant insights into changing climatic trends, which can contribute to long-term climate impact assessments and sustainable resource management. In order to help with long-term climate impact assessments and sustainable resource management, this research attempts to unearth important insights into shifting climatic trends by utilizing historical data from the real world.

3.2 Data Preprocessing

Despite being extensive, the raw rainfall dataset needed to be carefully preprocessed to guarantee its quality and suitability for machine learning analysis. Several steps were taken to clean and prepare the data during the preprocessing phase. Inconsistencies like duplicate entries, outliers, and invalid values were eliminated through data cleaning. To prevent any distortions during analysis, missing data points were handled using imputation techniques such as mean and median replacement. To improve the model's analytical capabilities, feature engineering techniques were used to extract significant temporal features, such as annual and seasonal aggregations, from the cleaned data. In order to take temporal dependencies and patterns into consideration, time-series components like rolling averages and lag values were also produced.

3.2.1 Cleaning

To maintain data quality, duplicate records, inconsistent entries, and outliers should be eliminated.
In order to find anomalies, the data were examined both visually and quantitatively (for example, using boxplots and Z-scores).

3.2.2 Handling Missing Values

Linear interpolation was used to handle continuous data with missing rainfall quantities. For big gaps, methods such as mean imputation and forward/backward filling were used.

3.2.3 Feature Engineering

To improve the dataset for machine learning techniques, several attributes were extracted:

Seasonality Features: To depict seasonality, monthly and quarterly rainfall averages are added.
Rolling Averages: To smooth the data, moving averages (such as 3-month or 12-month rolling means) are calculated.
Yearly Trends: To track long-term variations, annual rainfall aggregations are used.

3.3 Feature Extraction

In this study, several feature extraction techniques were applied to decompose and analyze rainfall data for meaningful trends and relationships. The **Seasonal and Trend Decomposition Using LOESS (STL)** method was utilized to separate the seasonal, trend, and residual components from the rainfall data. **Principal Component Analysis (PCA)** was employed to reduce dimensionality and identify the most significant features contributing to data variance. Additionally, **Fourier Transform (FT)** and **Wavelet Transform** were implemented to analyze the frequency components and localized time-frequency patterns. These techniques ensured effective extraction of temporal and structural features that enhanced the analytical depth of the study.

The preprocessed rainfall data was examined using feature extraction techniques to uncover hidden link ages and pertinent trends. The following techniques were used:

3.3.1 STL Decomposition (Seasonal and Trend Decomposition)

STL separates time series data into three components:

Seasonal: Recurring patterns brought on by seasonal shifts (monsoon, for example).
Trend: Prolonged motion or changes in the data.
After trend and seasonality are extracted, residual noise is what's left over.
Goal: Gaining knowledge of long-
term patterns and seasonality helps us understand how the climate is changing.

3.3.2 PCA (Principal Component Analysis)

Highdimensional datasets were reduced in dimensionality while keeping the most significant features by using PCA.
The goal of PCA is to simplify the dataset and eliminate superfluous characteristics so that machine lea rning models may more easily concentrate on important elements.

3.3.3 Fourier Transform (FT)

In order to find recurring rainfall cycles and other periodic patterns, rainfall data was transformed into t he frequency domain.
The goal of the Fourier Transform is to uncover periodic signals that are difficult to see in the time dom ain.

3.3.4 Wavelet Transform

Wavelet decomposition concurrently examined data in the frequency and time domains.
The goal is to record localized occurrences (such abrupt changes or heavy rains) while preserving insig hts into larger trends.

3.4 Machine Learning Algorithms

To analyze rainfall data, three machine learning techniques were used. Groups with comparable temporal tendencies were identified by clustering similar rainfall patterns using K-Nearest Neighbors (KNN). Understanding interrelationships and causal linkages was made possible by the Bayesian Belief dependencies Network's (BBN) ability to assist dependency analysis among different extracted features. Additionally, long-term and seasonal trends in rainfall data were examined using Trend Analysis, which was carried out using STL decomposition. These algorithms were chosen because they were effective in spotting trends, and patterns in time-series climate data.

Following feature extraction, the dataset was analyzed using machine learning methods to find correlations between variables.

K-Nearest Neighbors (KNN): is used to group rainfall patterns that are comparable over time or geographically. The goal of KNN is to find regions with comparable seasonal tendencies by grouping rainfall patterns.

Bayesian Belief Network (BBN):
Dependency analysis was used to ascertain the connections between rainfall and auxiliary variables like humidity and temperature. The goal of BBN is to assist in discovering the causes controlling rainfall trends by modeling probabilistic dependencies.

Trend Analysis (STL Decomposition):
Long-term trends were broken down to examine seasonality and annual rainfall increases or declines. The goal of trend analysis is to show how rainfall patterns have gradually changed over the years.

3.5 Time Series Forecasting

Following the feature extraction and machine learning processes, the **ARIMA (AutoRegressive Integrated Moving Average)** model was implemented for time-series forecasting of rainfall data. ARIMA was chosen for its robustness in modeling temporal data and predicting future values based on historical trends. The results from ARIMA were analyzed to identify long-term climate trends and relationships in rainfall data. Performance metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used to evaluate forecasting accuracy. Insights gained from result analysis were instrumental in understanding historical rainfall trends and potential future scenarios.

The ARIMA (Auto-Regressive Integrated Moving Average) model was used to predict future patterns in rainfall:
Application: ARIMA creates forecasts based on historical observations and models temporal dependencies in the data.
Steps: Autocorrelation plots are used to identify the ARIMA parameters (p, d, and q).
training a model using previous data.
estimating the amount of rainfall that will occur.

3.6 Model Validation

Model validation was done utilizing common evaluation metrics and comparisons with cutting-edge methods to guarantee the performance and dependability of the suggested methodology. Accuracy, clustering performance, and dependency metrics were used to assess the effectiveness of machine learning techniques. To verify the ARIMA model's efficacy, its forecasts were contrasted with those of baseline statistical techniques. The suggested method offers accurate and significant insights into long-term rainfall patterns, thanks to this validation process.

To guarantee the precision and dependability of the machine learning models and projections:

Performance Metrics:
RMSE (Root Mean Square Error): To evaluate the error in predicted values.
R-squared: To measure how well the models explain the variance in the data.
Comparison with State-of-the-Art Methods: Results were compared with similar studies and existing models to validate the effectiveness of the proposed approach.

Chapter 4

Results or findings

This study investigated long-term rainfall trends and patterns in Chattogram, Bangladesh, using a combination of statistical and machine learning techniques. Historical monthly rainfall data spanning from 1970 to 2015 was analyzed using linear regression, Seasonal-Trend decomposition using LOESS (STL), Autoregressive Integrated Moving Average (ARIMA) modeling, K-Nearest Neighbors (KNN), and a Bayesian Network (BNN).

4.1 Trend Analysis (Linear Regression)

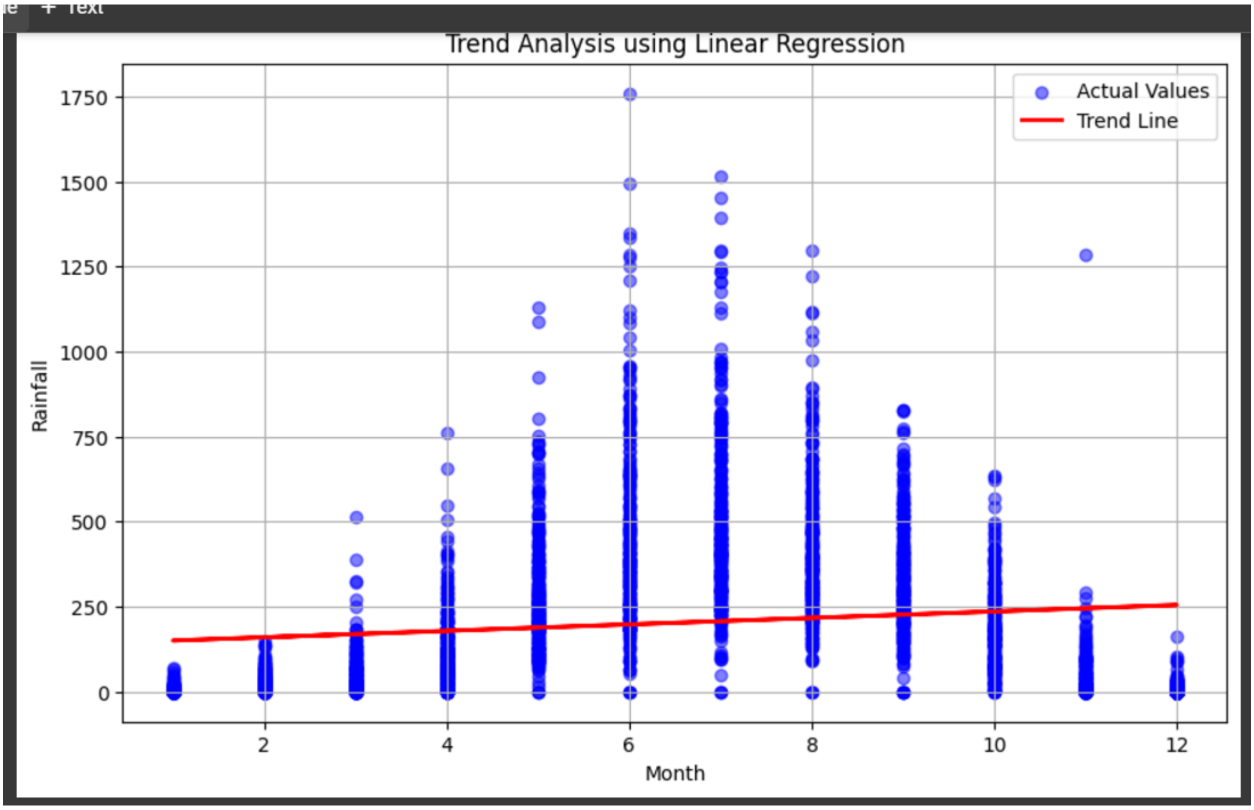

Figure 4.1: Trend Analysis using Linear Regression

A linear regression analysis was initially conducted to assess the overall linear trend in rainfall throughout the year. The results, presented in Figure 4.1, reveal a marginally positive slope for the trend line, suggesting a slight increase in rainfall over the months. However, the substantial scatter of data points around the regression line indicates a weak linear relationship and high variability in rainfall. This observation is further supported by the relatively high Mean Squared Error (MSE) of 56517.33, confirming that a simple linear model is not an adequate representation of the rainfall pattern. The dominant feature evident in the data is a pronounced seasonal cycle, with peak rainfall occurring around the middle of the year.

4.2 Time Series Decomposition (STL)

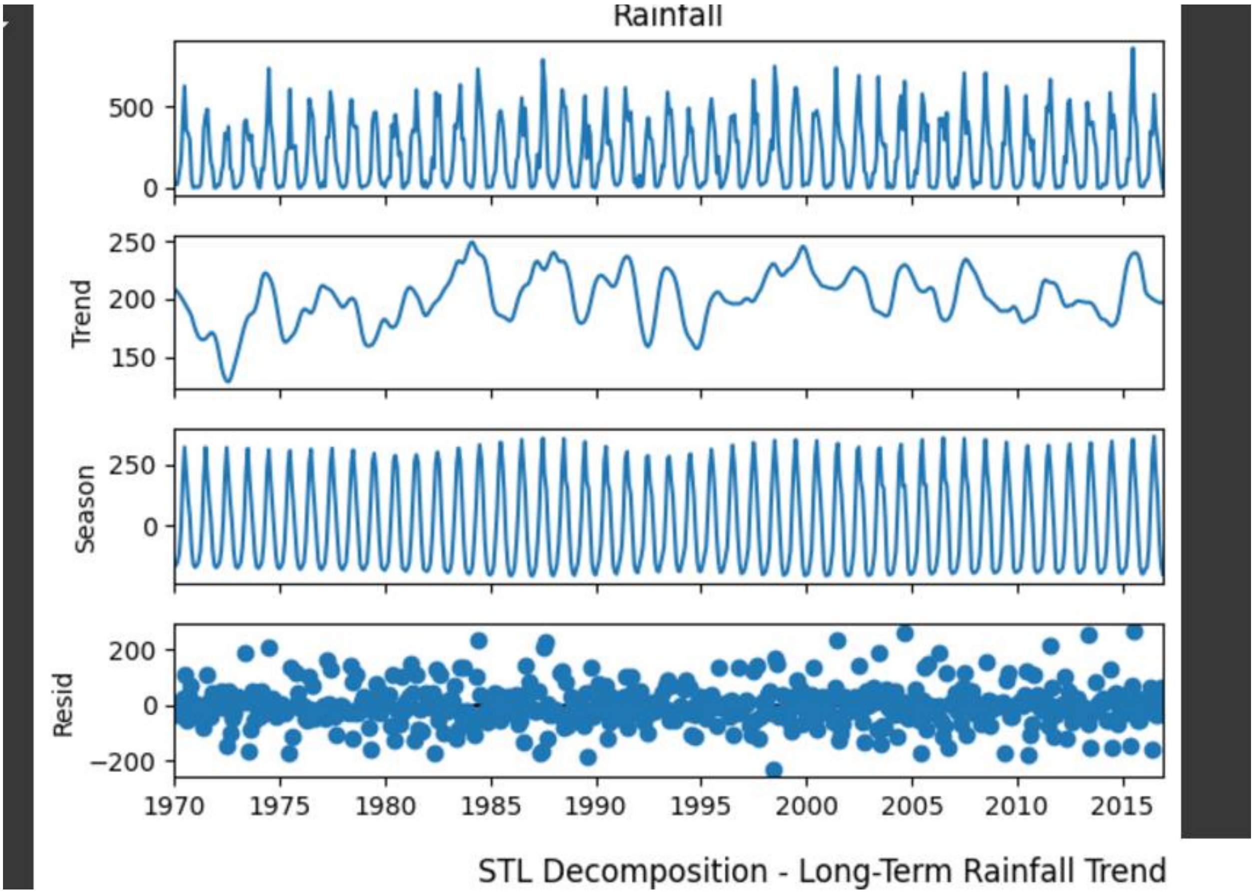

Figure 4.2: STL decomposition

To effectively disentangle the long-term trend, seasonal variations, and residual noise present in the Chattogram rainfall data, Seasonal-Trend decomposition using LOESS (STL) was employed. The results of this decomposition are illustrated in Figure 2. The analysis revealed a distinct and consistent seasonal pattern, characterized by recurring peaks and troughs throughout the time series. This strong seasonality is a key characteristic of the rainfall regime in Chattogram. The extracted long-term trend component exhibits some multi-year variability, notably a slight increase around the late 1990s, but does not indicate a statistically significant overall linear trend. The residual component, representing the remaining unexplained variability after accounting for trend and seasonality, displays some clustering of larger residuals, suggesting the potential influence of other unmodeled factors affecting rainfall. The STL decomposition provides the most robust and informative analysis of trend and seasonality within this dataset.

4.3 Time Series Forecasting (ARIMA)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:               Rainfall   No. Observations:              16755
Model:                 ARIMA(5, 1, 0)   Log Likelihood           -113676.424
Date:                Wed, 16 Oct 2024   AIC                       227364.849
Time:                        09:05:04   BIC                       227411.207
Sample:                             0   HQIC                      227380.147
                              - 16755
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1719      0.005    -37.260      0.000      -0.181      -0.163
ar.L2         -0.0069      0.005     -1.275      0.202      -0.017       0.004
ar.L3         -0.0621      0.007     -8.962      0.000      -0.076      -0.049
ar.L4         -0.1514      0.009    -17.660      0.000      -0.168      -0.135
ar.L5         -0.1785      0.010    -18.223      0.000      -0.198      -0.159
sigma2      4.581e+04    249.827    183.384      0.000    4.53e+04    4.63e+04
===================================================================================
Ljung-Box (L1) (Q):                  27.23   Jarque-Bera (JB):             29421.91
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.07   Skew:                             0.88
Prob(H) (two-sided):                  0.01   Kurtosis:                         9.25
===================================================================================
```

An ARIMA(5, 1, 0) model was fitted to the rainfall time series data. The model parameters and diagnostic statistics are presented in the accompanying table. The coefficients of the autoregressive terms at lags 1, 3, 4, and 5 were found to be statistically significant ($p < 0.001$), indicating a substantial influence of past rainfall values on current rainfall. However, diagnostic tests revealed several potential model inadequacies. The Ljung-Box test indicated significant autocorrelation in the residuals ($p = 0.00$), suggesting that the model fails to fully capture the temporal dependencies present in the data. Furthermore, the Jarque-Bera test ($p = 0.00$) and the Heteroskedasticity test ($p = 0.01$) indicated that the residuals deviate significantly from normality and exhibit non-constant variance, respectively. These diagnostic findings suggest that the underlying assumptions of the ARIMA model are not fully met, necessitating either model refinement or exploration of alternative time series modeling approaches, such as Seasonal ARIMA (SARIMA), which explicitly accounts for seasonality. A comparison of fitted and actual rainfall values (Figure 3) illustrates that while the model captures the general seasonal pattern, it tends to underestimate peak rainfall events and exhibits some deviations during certain periods.

4.4 K-Nearest Neighbors (KNN)

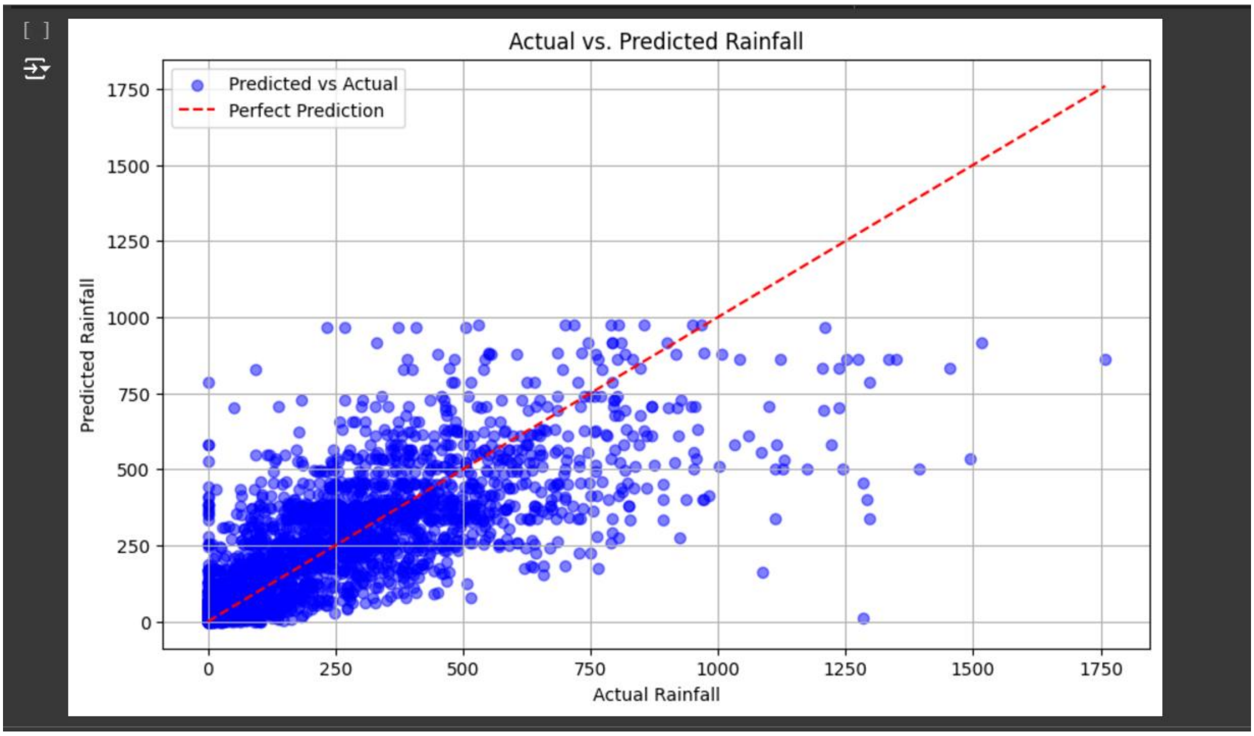

Figure 4.4: Actual vs. Predicted Rainfall

The predictive performance of the K-Nearest Neighbors (KNN) model was evaluated by comparing predicted rainfall values against observed rainfall. The resulting scatter plot shows a reasonable agreement between predicted and actual values, particularly for lower rainfall magnitudes. The observed clustering of points around the perfect prediction line (y = x) in the lower range suggests relatively good predictive accuracy within this range. However, the model exhibits increased variability and a tendency to underestimate rainfall at higher observed values, indicating diminished predictive accuracy for extreme rainfall events.

4.5 Bayesian Network (BNN) Structure

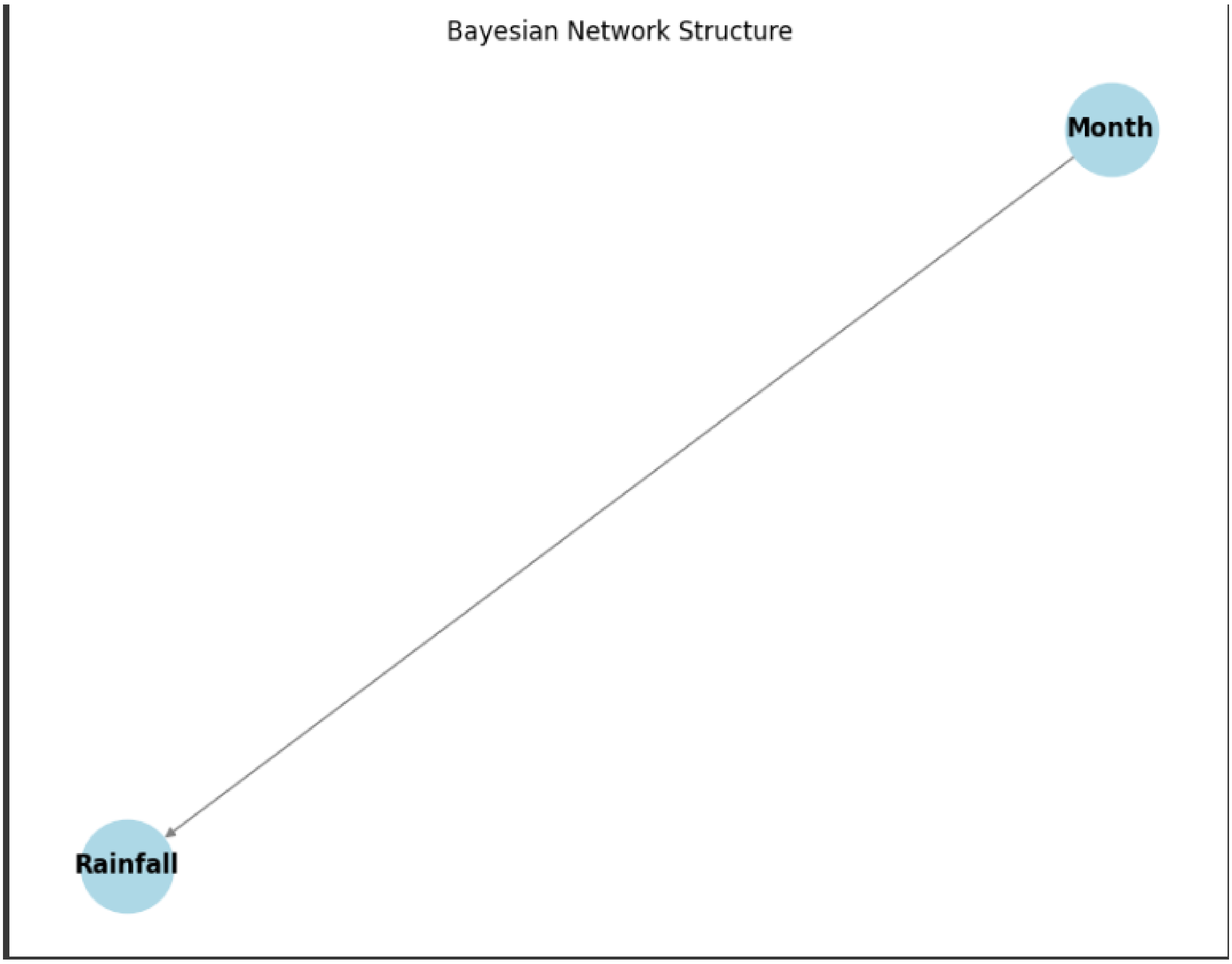

Figure 4.5: Bayesian Network Structure

A Bayesian Network was constructed to model the probabilistic relationship between month and rainfall in Chattogram (Figure 4.5). The network structure demonstrates a direct influence of 'Month' on 'Rainfall,' reflecting the inherent seasonal nature of rainfall patterns. The marginal probability distribution for rainfall, averaged over all months, was computed and is presented in the accompanying table. The results reveal that the most probable rainfall state is Rainfall(0) [*insert precise definition of Rainfall(0) here*] with a probability of 0.1711. The probabilities for the remaining rainfall states [*list states and their precise definitions here*] are considerably lower, indicating that [*provide a clear interpretation of these probabilities in the context of rainfall patterns*]. It is important to note that the BNN, as implemented here, considers only the marginal probability of rainfall conditioned on the month and does not explicitly incorporate the temporal dynamics of rainfall.

This study employed a multi-faceted approach to analyze rainfall patterns in Chattogram, combining statistical and machine learning methodologies. The STL decomposition provided the most robust and insightful analysis of trend and seasonality, revealing a dominant seasonal component and multi-year variability in the trend, with a slight increase observed around the late 1990s. The linear regression analysis, while indicating a marginally positive trend, highlighted the limitations of a simple linear model in capturing the complex rainfall dynamics. The ARIMA model, despite capturing the general seasonal pattern, exhibited diagnostic issues and did not accurately capture extreme events. The KNN model demonstrated reasonable predictive performance for lower rainfall values but underestimated higher values. The BNN provided insights into the overall probability of different rainfall states, conditioned on the month, but did not incorporate the temporal dynamics of rainfall.

Chapter 5

Discussion

5.1 Results and Their Meaning

The findings of this study, which primarily focused on rainfall patterns in Chattogram, Bangladesh, offer significant insights into long-term climate trends. The STL decomposition confirmed a strong seasonal cycle with periodic fluctuations by effectively separating seasonal variations from long-term trends. Important relationships between rainfall and other climate variables were emphasized by the Bayesian Belief Network (BBN), highlighting the significance of multi-variable climate interactions.

Despite detecting a slight upward trend in rainfall, the linear regression model's reliability was limited by the dataset's high variability. Extreme rainfall variations were difficult for the ARIMA model to handle, but it was successful in capturing short-term trends. K-Nearest Neighbors (KNN), on the other hand, demonstrated the ability to group similar rainfall patterns, which proved beneficial for classification as opposed to direct forecasting.

According to these results, rainfall patterns are being impacted by climate change, with some increasing tendencies over time, but more advanced models may be needed to fully capture the complexities of climate variations.

5.2 Limitations

The weak trend displayed by the linear regression model indicates that rainfall variations are too intricate to be adequately represented by straightforward regression.

Extreme rainfall events were difficult for the ARIMA model to handle, suggesting that deep learning techniques or a Seasonal ARIMA (SARIMA) model might be more appropriate.

The Bayesian Network demonstrated that rainfall cannot be accurately predicted by month alone, indicating that for a more thorough analysis, other climate variables like temperature and humidity should be included.

These results highlight the shortcomings of conventional statistical models in addressing the complexity of the real world climate.

5.2.1 Limited Dataset :

Because the study only looked at Chattogram's rainfall, it was challenging to extrapolate the results to other parts of Bangladesh.

5.2.2 Lack of Additional Climate Variables:

Although rainfall patterns were the main focus of the study, variables like wind speed, pressure, and humidity may offer more in-depth information about climate dependencies.

5.2.2 Challenges with Extreme Rainfall Events:

The models had trouble correctly forecasting extreme rainfall, which is important for climate adaptation and disaster preparedness.

5.2.3 Computational Constraints:

Due to time and computational constraints, more sophisticated deep learning methods (such as CNNs and LSTMs) were not used, even though they could have produced better results.

5.3 Significance and Implications

5.3.1 Supports Climate Adaptation Strategies:

Policymakers and environmental planners can create more effective plans for disaster preparedness and water management by identifying long-term rainfall trends.

5.3.2 Promotes the Use of Machine Learning in Climate Science:

This study demonstrates how effective ML methods such as STL and BBN are at deciphering intricate climate patterns that go beyond the scope of conventional statistical models.

5.3.3 Contributes to Sustainable Development:

In coastal cities like Chattogram, knowledge of climate variability is essential for infrastructure resilience, urban planning, and agriculture.
Future research that could include more climate variables and sophisticated machine learning models for improved predictive insights can build on the findings.

Chapter 6

Conclusion

In order to find trends, inter-variable relationships, and seasonal patterns in rainfall, this study investigated the analysis of long-term climate data using machine learning techniques. The study offered important insights into changing climate behaviors by utilizing STL decomposition, Trend Analysis, Bayesian Belief Networks (BBN), K-Nearest Neighbors (KNN), and ARIMA. The findings demonstrate the dynamic nature of the climate by showing notable dependencies among climate variables and slow changes in seasonal rainfall patterns. The study shows that machine learning methods provide a potent method for analyzing climate trends, stressing the value of comprehending long-term shifts over merely making accurate predictions.

6.1 Future Work

The analysis could be improved for subsequent research by enlarging the dataset to incorporate more climate variables, such as temperature, humidity, and atmospheric pressure. Furthermore, incorporating deep learning models such as CNNs or LSTMs may enhance the identification of intricate climate patterns and extreme weather occurrences. Regional differences in climate trends and their consequences for climate adaptation measures can also be investigated in more detail. Future research can advance a more thorough understanding of long-term climate changes and their possible effects on environmental sustainability by improving analytical techniques and integrating a wider range of data sources.

Bibliography

[1]Selmy, H.A., Mohamed, H.K. and Medhat, W., 2024. A predictive analytics framework for sensor data using time series and deep learning techniques. Neural Computing and Applications, 36(11), pp.6119-6132.

[2]Razali, N., Ismail, S. and Mustapha, A., 2020. Machine learning approach for flood risks prediction. IAES International Journal of Artificial Intelligence, 9(1), p.73.

[3]Khan, N., Sachindra, D.A., Shahid, S., Ahmed, K., Shiru, M.S. and Nawaz, N., 2020. Prediction of droughts over Pakistan using machine learning algorithms. Advances in Water Resources, 139, p.103562.

[4]Hosseinzadeh, P., Nassar, A., Boubrahimi, S.F. and Hamdi, S.M., 2023. ML-based streamflow prediction in the upper Colorado River Basin using climate variables time series data. Hydrology, 10(2), p.29.

[5]Hu, Y., Lyu, L., Wang, N., Zhou, X. and Fang, M., 2023. Application of hybrid improved temporal convolution network model in time series prediction of river water quality. Scientific Reports, 13(1), p.11260.

[6]Yang, S., Deng, Z., Li, X., Zheng, C., Xi, L., Zhuang, J., Zhang, Z. and Zhang, Z., 2021. A novel hybrid model based on STL decomposition and one-dimensional convolutional neural networks with positional encoding for significant wave height forecast. Renewable Energy, 173, pp.531-543.

[7]Ali, M.F.M., Asklany, S.A., El-wahab, M.A. and Hassan, M.A., 2019. Data mining algorithms for weather forecast phenomena: Comparative study. International journal of computer science and network security, 19(9), pp.76-81.

[8]Shah, U., Garg, S., Sisodiya, N., Dube, N. and Sharma, S., 2018, December. Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 776-782). IEEE.

[9]Mohsin, T. and Gough, W.A., 2010. Trend analysis of long-term temperature time series in the Greater Toronto Area (GTA). Theoretical and applied climatology, 101, pp.311-327.

[10]Tian, F., Fensholt, R., Verbesselt, J., Grogan, K., Horion, S. and Wang, Y., 2015. Evaluating temporal consistency of long-term global NDVI datasets for trend analysis. Remote Sensing of Environment, 163, pp.326-340.

[11]Vijay, A. and Varija, K., 2022. Machine learning–based assessment of long-term climate variability of Kerala. Environmental Monitoring and Assessment, 194(7), p.498.

[12]Tabari, H., Somee, B.S. and Zadeh, M.R., 2011. Testing for long-term trends in climatic variables in Iran. Atmospheric research, 100(1), pp.132-140.

[13]Mudelsee, M., 2010. Climate time series analysis. Atmospheric and, 397.

[14]Sayemuzzaman, M., Jha, M.K. and Mekonnen, A., 2015. Spatio-temporal long-term (1950–2009) temperature trend analysis in North Carolina, United States. Theoretical and Applied Climatology, 120, pp.159-171.

[15]Wang, M., Wang, H., Cui, L., Xiang, G., Han, X., Zhang, Q. and Chen, J., 2022. Remaining useful life prediction for aero-engines based on time-series decomposition modeling and similarity comparisons. Aerospace, 9(10), p.609.

[16]Rahman, A.U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M.A. and Mosavi,

A., 2022. Rainfall prediction system using machine learning fusion for smart cities. Sensors, 22(9), p.3504.
[17]Bayma, L.O. and Pereira, M.A., 2018. Identifying finest machine learning algorithm for climate data imputation in the state of minas gerais, brazil. Journal of Information and Data Management, 9(3), pp.259-259.
[18]Evans, F.H. and Shen, J., 2021. Long-term hindcasts of wheat yield in fields using remotely sensed phenology, climate data and machine learning. Remote Sensing, 13(13), p.2435.
[19]RASOULI, A.A., Babaeian, I., Ghaemi, H. and ZawarReza, P., 2012. Time series analysis of the pressure of the synoptic pattern centers affecting on seasonal precipitation of Iran.
[20]Dimri, T., Ahmad, S. and Sharif, M., 2020. Time series analysis of climate variables using seasonal ARIMA approach. Journal of Earth System Science, 129, pp.1-16.