REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

PREDICTION OF WEATHER USING DATA MINING TECHNIQUES

Abdulqader Mohammed Hashim AL-ANI

Master's Thesis

Supervisor

Asst. Prof. Dr. Sefer KURNAZ

Istanbul,2023

PREDICTION OF WEATHER USING DATA MINING TECHNIQUES

Abdulqader Mohammed Hashim AL-ANI

Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

PREDICTION OF WEATHER USING DATA MINING TECHNIQUES prepared by ABDULQADER MOHAMMED HASHIM AL-ANI and submitted on 12/April /2023 has been **accepted unanimously/by majority of votes** for the degree of Master in INFORMATION TECHNOLOGY

Academic Title – First and Last Name of the Co-Supervisor	Academic Title – First and Last Name of the Supervisor

Thesis Defense Committee Members:

Academic Title - First/Last Name	Department, University	____________________
Academic Title - First/Last Name	Department, University	____________________
Academic Title - First/Last Name	Department, University	____________________
Academic Title - First/Last Name	Department, University	____________________
Academic Title - First/Last Name	Department, University	____________________

Accepted by Altınbaş University Institute of Graduate Studies on the following date:

____/____/____.

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Abdulqader Mohammed Hashim AL-ANI

Signature

PREFACE

I might want to thank my administrator: Supervisor Name Please let me express. my profound feeling of appreciation and gratefulness to both of you for the information: direction and unrestricted help you have given me. I want you to enjoy all that life has to offer. and further achievements and accomplishments throughout your life.

ABSTRACT

PREDICTION OF WEATHER USING DATA MINING TECHNIQUES

AL-ANI, Abdulqader Mohammed

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: April / 2023

Pages: 65

Weather forecasting has become increasingly essential as a result of its uses in a variety of industries, including agriculture, energy companies, and everyday life. Weather prediction has become a real-time challenge for the world in the previous decade. Because of the ever-changing meteorological conditions, forecasting is becoming more difficult. Weather forecasting is the prediction of how the current state of the atmosphere will change in the future. Understanding the numerous contributing elements that produce weather variations is crucial for efficient weather analysis. The process of recording meteorological data such as wind direction, wind speed, humidity, rainfall, temperature, and so on is known as weather forecasting. Since machine learning techniques are more robust to perturbations, in this project we applied linear regression and LSTM to predict the weather such as temperature, rainfall etc. and compare both approaches and analyzed it. We used two different datasets for the same. Coming to result that we got from each approaches was quite amazing. In the linear regression approach, we got mean absolute error about 96.32 mm and 2.69 celsius when performing rainfall and temperature prediction respectively whereas in the deep learning approach, the mean absolute error was 0.002268 degree celsius, 0.003266 km/h, and 0.003069 Pascal when performing temperature, wind speed and pressure prediction respectively. We could clearly see the difference between the outcomes.

Keywords: Weather Prediction, Deep Learning, Machine Learning, SRIMA, Personal Computer.

TABLE OF CONTENTS

Pages

ABSTRACT .... **viii**

LIST OF FIGURES .... **xii**

LIST OF TABLES .... **xiv**

ABBREVIATIONS .... **xv**

1. INTRODUCTION **1**

1.1 INTRODUCTION 1

1.2 PROBLEM STATEMENT 2

1.3 RESEARCH OBJECTIVE 2

1.4 SIGNIFICANCE OF WEATHER FORECASTING 3

1.5 CONTRIBUTION 4

2. BACKGROUND **6**

2.1 WEATHER DATA IDENTIFICATION 6

2.2 WEATHER FORECAST METHODS 7

2.3 LITERATURE REVIEW 9

2.4 RELATED WORKS 10

2.5 CHAPTER SUMMARY 14

3. PROPOSED SYSTEM **15**

3.1 METHEDOLOGY 15

3.1.1 Data Collection 17

3.1.1.1 Data preprocessing 18

3.1.1.2 Feature extraction 19

3.1.2 Regression 20

3.1.2.1 Linear regression 20

3.1.2.2 Decision tree 21

3.1.3 Classification 22

3.1.3.1 Artificial neural networks (ANN) 22
3.1.3.2 Lstm 23
3.1.3.3 Forget gate 24
3.1.3.4 Input gate 25
3.1.3.5 Output gate 25
3.1.3.6 Cell state 26
3.1.3.7 Conventional neural network 26
3.2 CONCLUSION 27
4. RESULT AND DISCUSSION 28
4.1 EXPERIMENT SETUP 28
4.2 DATA ANALYSIS 29
4.3 PRINCIPAL COMPONENT ANALYSIS (PCA) 33
4.4 REGRESSION RESULTS 34
4.4.1 Linear Regression Results 34
4.4.2 Decision Tree 37
4.5 CLASSIFICATION RESULTS 39
4.5.1 Cyclone Detection by Convolutional Neural Network (CNN) 40
4.5.2 Monsoon Classification by LSTM and Seasonal ARIMA Model 43
4.5.3 Recognizing the Month by Outgoing Longwave Radiation (OLR) 45
4.5.4 Retrieving Winds From the Vorticity and Divergence Fields Using ANN 48
4.6 CONCLUSION 50
5. CONCLUSION AND FUTURE WORKS 51
5.1 CONCLUSIONS 51
5.2 FUTURE WORKS 51
REFERENCES 52

LIST OF FIGURES

Pages

Figure 2.1: An Overview of Situations Employing Machine Learning in Tropical Cyclone Forecasting. 13

Figure 3.1: Methodology Block Diagram 18

Figure 3.2: Decision Tree 21

Figure 3.3: Neural Network Architecture Diagram 23

Figure 3.4: LSTM Architecture 24

Figure 4.1: Extracting Features 30

Figure 4.2: The Frequency of Data Distribution 31

Figure 4.3: Pair plots for T_mu, P_mu, Td_mu, Ff_mu, VV_mu, U_mu 32

Figure 4.4: The Correlation Matrix 33

Figure 4.5: The 3 Components 34

Figure 4.6: Linear Regression Results 35

Figure 4.7:Linear Regression Results with PCA 36

Figure 4.8: Decision Tree Model 38

Figure 4.9: The Mean Sea Level Pressure 41

Figure 4.10: CNN Classification Results 42

Figure 4.11: Confusion Matrix 42

Figure 4.12: SRIMA Model diagnostics 44

Figure 4.13: SRIMA Classification Results 44

Figure 4.14: LSTM Classification Results 45

Figure 4.15: OLR Data for the 15th of Every Month 46

Figure 4.16: Test Data Has an Accuracy. 47

Figure 4.17: Confusion Matrix 47

Figure 4.18: Training and Validation Losses 49

Figure 4.19: Predicted Outputs. 49

LIST OF TABLES

Pages

Table 3.1: Dataset Description 17

Table 4.1: Model Performance 35

Table 4.2: Decision Tree Prediction Results 38

ABBREVIATIONS

CNN	:	Conventional Neural Network
SUN	:	Scene Understanding
URDF	:	Unified Robot Description Format
PC	:	Personal Computer
DRL	:	Deep Reinforcement Learning
MDP	:	Markov Decision Process
SLAM	:	Simultaneous Localization and Mapping Problem

1. INTRODUCTION

This part will take a gander at how the 'Web of Things' and 'Huge Information' are being coordinated in the disciplines of data innovation and business. Each plan was assessed in light of its effect and potential for savvy connected networks. This part looks at how these advancements have upset the world by recognizing and examining a few potential open doors and moves introduced by the capacity to ingest and use monstrous measures of 'Web of Things' information, remembering applications for shrewd urban communities, savvy horticulture, brilliant transportation, etc. At long last, this part examines how the assortment of ongoing information from 'Web of Things' gadgets offered an opportunity to utilize information mining strategies and fabricate brilliant weather conditions figures.

1.1 INTRODUCTION

Metropolitan people group are supposed to confront various difficulties soon, including health, legitimacy, energy usage, compelling assistance movement, strong transportation systems, etc. Progresses in network coordination of information systems, recognizing and concentrated gadgets, data sources, and man-made mental ability are empowering the improvement of wise and connected networks. Individuals these days request the force of sensors, appropriated processing, fast associations, and data investigation when they utilize different devices like PDAs, cars, relational associations, and transportation applications like Uber.

Insights gauge that 500 billion devices will be associated with the Web by 2030, as per [46]. With the assistance of inserted innovation, every contraption has sensors that assemble information, associate with the climate, and impart across the organization with practically no administrator inclusion. The organization of these connected gadgets is known as the Web of Things (IoT).

On September 14, 2015, the US government reported another Brilliant Urban area Drive to help nearby networks in tending to significant worries, for example, bringing down gridlock, fighting wrongdoing, animating financial development, controlling the results of environmental change, and further developing city administration conveyance. The Systems administration and Data Innovation Innovative work (NITRD) Program reported the arrival of rendition 4 of a Shrewd and Associated People group Structure on November 25, 2015. (NITRD, 2015). The essential reason is that networks in all settings and scales approach Web of Things advancements and administrations to work on their residents' wellbeing, security, and economy.

1.2 PROBLEM STATEMENT

In this thesis, we assembled a weather prediction model that successfully anticipate the weather. Our model is concerned with meteorological data in the form of time series, and those data are changed with information where meaningful information is derived using data mining techniques. The data mining system is used to uncover hidden patterns and connect them to the many parameters associated with meteorological conditions. The data now displays a series across time, and weather prediction can be well considered by applying time series mining. Several automated prediction systems on the market are being investigated for providing incorrect information to customers. Several prediction systems on the market lack capabilities and appropriate interface designs. As a result, individuals believe that all prediction systems are the same. To eliminate these misconceptions, our technology will deliver appropriate degrees of prediction while occupants are away from their home. This system will be especially beneficial to people who work irregular hours at their jobs and take frequent vacations. Eventually, a flawed prediction system is destructive to people's assets; it has turned into a drama with people's trust. As a result, we see it as a possibility for working on this platform.

1.3 RESEARCH OBJECTIVE

The main objectives of the project on weather classification and prediction using Linear Regression, KNN, Decision Tree, CNN, and LSTM are:

a. To explore the use of various machine learning techniques for weather classification and prediction.

b. To preprocess and extract features from weather data to be used as inputs for the different models.

c. To implement and compare the performance of Linear Regression, KNN, Decision Tree, CNN, and LSTM models for weather classification and prediction tasks.

d. To evaluate the accuracy and efficiency of each model and identify the strengths and limitations of each technique.

e. To identify the factors that affect the accuracy and reliability of the models, such as the quality and representativeness of the input data.

f. To determine the optimal model or combination of models for different weather prediction scenarios based on the evaluation results.

g. To demonstrate the practical applications of machine learning techniques in weather prediction and highlight the potential benefits for various industries and sectors..

1.4 SIGNIFICANCE OF WEATHER FORECASTING

Weather conditions estimating is quite possibly of the main subject that researchers and specialists have been moved with because of an anticipated expansion in climate occasions and environment switches up the world. Weather conditions determining is the utilization of science and innovation to conjecture the condition of the air at a particular region. Since old times, people have tried to foresee the climate, and logical estimate models have been created since the eighteenth hundred years. Forecasters currently gather meteorological information utilizing more present day systems and innovation, as well as the world's most remarkable PCs. The fresh debuts are further developing in data gathering limits, with the capacity to send off satellites and supercomputers and reap information from IoT gadgets. Moreover, the utilization of information examination procedures, as well as AI, man-made brainpower, and cloud-based cautioning frameworks; for instance, a framework that lets a carrier know when to reschedule trips to stay away from rainstorms, or a rancher when to water a particular line of harvests. These models can be modified to figure changes in the environment and climate. Regardless of these headways, weather conditions gauges are still regularly mistaken. Since weather conditions is a confounded and erratic framework, it is unbelievably challenging to expect.

Weather conditions anticipating benefits numerous parts of society's social and monetary prosperity. Weather conditions guaging is basic data for some ventures, including horticulture, avionics, energy, trade, oceanic, warnings, protection firms, etc. It can likewise considerably affect direction and regulation, worldwide food security, development arranging, efficiency, and hazard the board in the climate [1].

Weather conditions gauging is often used to foresee and caution about cataclysmic events brought about by unexpected changes in weather conditions. Outrageous weather conditions are turning into a rising obligation to the economy, attributable to environmental change, with around ten climate and environment fiascos costing more than $1 billion each as of late, as per NOAA. Western rapidly spreading fires have cost more than $40 billion over the most recent two years alone. Tropical storms are pouring more downpour than previously, while heat waves are turning out to be all the more remarkable and successive (Freedman, 2019).

The Public Weather Conditions Administration is as of now in an abnormal circumstance as it endeavors to achieve its motivation of safeguarding lives and property. With regards to climate and environment perceptions, the office is being tried on the grounds that a few confidential associations are obtaining unrivaled information. As indicated by a 2017 Public Weather Conditions Administration report, the confidential weather conditions determining industry is valued at $7 billion (and developing). It's additionally putting the national government's hold on climate information and alerts to the test. [2].

1.5 CONTRIBUTION

Weather classification and prediction are essential machine learning applications in the realm of weather forecasting. Weather categorization and prediction techniques include Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. Here's how each method can be used to classify and predict weather:

Linear Regression: Straight Relapse is an essential calculation that can be utilized to foresee the climate. Direct relapse can be utilized to gauge the connection between climate factors like temperature, moistness, and wind speed and the objective variable like precipitation or snowfall. In the wake of preparing the model with past information, you can utilize it to conjecture future atmospheric conditions. Direct Relapse is a basic and clear methodology for climate expectation. It is quick to prepare and computationally proficient. Be that as it may, it suggests a direct connection between the info factors and the objective variable, which in weather conditions estimating may not generally be the situation.

K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm that does not make any assumptions about the underlying distribution of the data. It is useful in cases where the relationship between the input variables and the target variable is complex and non-linear. However, KNN can be computationally expensive, especially when dealing with large datasets.

Decision Tree: The Choice Tree calculation is a notable method that can be utilized for both climate order and expectation. It is easy to utilize and upholds both classification and constant info factors. The Choice Tree can likewise be utilized to track down key components that guide in climate grouping and expectation. Sadly, Choice Tree is vulnerable to overfitting, especially while managing boisterous information.

Convolutional Neural Networks (CNN): CNN is a deep learning algorithm that can be used for weather classification. It can automatically learn relevant features from the input data, making it useful for image-based weather classification. CNN has achieved state-of-the-art performance in various image classification tasks, including weather classification. However, CNN requires large amounts of training data and can be computationally expensive, especially when training on large datasets.

Long Short-Term Memory (LSTM): The LSTM is a kind of intermittent brain network that can be utilized to foresee climate. It is significant for time-series climate expectation since it can catch long haul conditions in the information. LSTM has shown state of the art execution in an assortment of time-series forecast errands, including climate expectation. LSTM, then again, takes a significant amount of preparing information and can be computationally costly, especially while preparing on lengthy time series information.

In conclusion, Linear Regression, KNN, Decision Tree, CNN, and LSTM are all algorithms that can be used for weather classification and prediction. The choice of algorithm depends on the nature of the problem, the quality of the data, and the level of accuracy required. It is often useful to compare the performance of different algorithms using metrics such as accuracy, precision, recall, and F1 score.

2. BACKGROUND

This part will cover the accessible writing by different scientists who have made and utilized calculations in view of information digging approaches for climate expectation. Also, various expectation logical methodologies are explored and examined regarding different factors and patterns. There are different order and expectation calculations. K-implies bunching, Choice Trees, Counterfeit Brain Organizations, Backing Vector Machines (SVM), Bayesian Order, and Relapse are a couple of models. There are different rules for assessing a calculation's expectation execution. This scientists' work has been analyzed and looked at in a table organization.

2.1 WEATHER DATA IDENTIFICATION

Weather conditions determining starts with a steady perception of the situation with the air for a specific site and its environmental elements. Temperature, gaseous tension, dampness, precipitation, dissipation, mugginess, wind speed, wind course, and other barometrical elements should be checked and recorded. The World Meteorological Association offers the establishment for noticing stages like radars, satellites, and surface climate perceptions, which help in the observing of different climate factors. The Public Maritime and Environmental Organization (NOAA) is the sole approved backer of serious weather conditions watches and admonitions in the US, and it is generally viewed as the forerunner in precise weather conditions estimates and lifesaving alerts.

Weather conditions stations incorporate a thermometer and an indicator. Different sensors evaluate environmental factors, for example, wind speed, wind heading, dampness, and precipitation sum. This hardware is put at different areas to evaluate the air boundaries of that area. Climate information is gathered by 15 satellites, 100 fixed floats, 600 floating floats, 3,000 planes, 7,300 boats, and around 10,000 land-based stations, as indicated by the WMO. The Mechanized Surface Noticing Framework is the name given to the authority weather conditions stations used by the Public Weather conditions Administration (ASOS). As they fly through the air, radiosondes measure environmental properties like temperature, tension, and stickiness. Wind speed and heading can be gotten by following radiosondes in flight.

Radar is an abbreviation that represents Radio ID and Going. A transmitter conveys radio waves that bob off neighboring items and return to a recipient. Environment radar can separate a couple of parts of precipitation, like its area, speed, seriousness, and probability of future precipitation. Doppler radar can likewise identify how rapidly precipitation falls. Radar can approach the development of a tempest and foresee its logical outcomes. Starting from the first was sent off in 1952, weather conditions satellites have become progressively fundamental suppliers of meteorological information, and they are the best means to screen enormous scope cycles like tempests. Weather conditions conjectures are regularly directed at the large-scale level by gathering information from remote detecting satellites [3]. Satellites might gather photographs that undertaking weather patterns, for example, greatest and least temperatures, precipitation degree, cloud conditions, wind streams, and their headings. Satellite-based frameworks are intrinsically more costly and need a thorough help framework. Moreover, such frameworks can give such data, which is normally summed up over a more prominent topographical region.

Moreover, satellites can record long haul changes, for example, how much ice cover over the Cold Sea every September. They additionally see all frequencies of radiation in the electromagnetic range. The Public Weather Conditions Administration depends on information from the Geostationary Functional Ecological Satellites (GOES). GOES are grouped into three kinds: visual, infrared, and water fume. Noticeable makes light visuals of tempests, mists, flames, and exhaust cloud. Mists, ocean and land temperatures, and sea peculiarities, for example, sea flows are undeniably recorded utilizing infrared symbolism. Water fume symbolism is the last type of GOES symbolism. This imaging looks at the dampness content in the upper part of the sky to decide if fogs can rise to gigantic levels. One more kind of satellite that is generally utilized in weather conditions determining is the Polar Circling Natural Satellite (POES). These satellites fly a lot nearer to Earth, around 530 miles away, and circle the planet from one shaft to another.

2.2 WEATHER FORECAST METHODS

Individuals can be made aware of outrageous climate or changes in air factors, for example, temperature, wind speed and course, stickiness, daylight, shadiness, and precipitation, on account of new mechanical leap forwards and information examination strategies. Weather conditions is depicted with regards to the condition of the environment at a specific area and time by deviations from these models. Weather conditions conjectures ordinarily follow three key cycles, as per [4]:

a. Atmosphere, ocean, and land surface observation and data collecting.

b. Osmosis, handling, investigation, and extrapolation to expect the future condition of the air.

c. A perception framework, examination, a model, and a PC framework include a conjecture framework. Brief weather conditions estimating, mathematical systems, and measurable strategies are instances of weather conditions gauging techniques.

a. The exemplary strategy to climate forecast is succinct climate expectation. Weather conditions estimates are based on succinct graphs. Brief graphs depend on gigantic measures of observational information of different climate factors got from weather conditions stations at a specific period. Consistently, meteorological establishments make a progression of succinct diagrams to follow weather conditions changes, which are the groundwork of weather conditions conjectures. All through numerous long periods of cautious investigation of weather conditions diagrams, a few observational regulations were created. These characterized rules have further developed gauges assessing the speed and course of climate framework movement.

b. Mathematical climate expectation (NWP) is perhaps of the most basic functional work performed by meteorological administrations overall [5]. NWP estimates weather conditions utilizing PC calculations. Supercomputers execute muddled PC frameworks that expect different environmental circumstances. This procedure utilizes a bunch of halfway differential conditions and different details to portray the dynamic and thermodynamic cycles that happen in the world's environment. It incorporates conditions, mathematical approximations, parametrizations, area settings, and beginning and limit conditions.

c. Along with mathematical methodologies, factual climate expectation is utilized. This technique depends on climate information from an earlier time and searches for components that are great indicators of future occasions. The essential objective is to figure out what parts of weather conditions are great indicators of future occasions. With the foundation of these linkages, precise information can be dependably used to estimate future circumstances.

This technique can anticipate the general climate. The factors characterizing weather patterns, like temperature (greatest or least), relative mugginess, precipitation, etc, differ constantly over the long run, framing time series of every boundary and can be utilized to foster an estimating model measurably or through different means, like Fake Brain Organization (ANN) and Choice Tree calculations.

2.3 LITERATURE REVIEW

There are various information examination approaches and factual models for weather conditions determining in the writing [6]. Various experts have applied and made different assumption models either really or by using one or two technique, for example, relapses, brain organizations, choice trees, grouping, and different information mining procedures trying to figure out the best genuine opportunity information examination strategy to apply on weather conditions estimating [7]. To foresee precipitation, the review utilizes least and most extreme temperatures, relative mugginess, and wind speed as information boundaries. To stay away from overfitting, information was partitioned into two sections: 70% for preparing and 30% for testing. Choice trees, arbitrary timberlands, K-closest neighbor, brain organizations, and Backing Vector Relapse were the AI calculations utilized on the information (SVR). The forecast procedures were analyzed and picked in light of the Root Mean Square Blunder (RMSE), while the AI strategy was thought about utilizing the Region Under Bend (AUC) (AUC). These assessments recommend that ARIMA is the best methodology for anticipating temperatures, SVR is the most ideal way for anticipating mugginess and wind speed, and the irregular woodland strategy predicts precipitation with the most elevated exactness.

[8] proposed a nonexclusive steady K-mean grouping calculation for weather conditions determining. The creators represented how device grouping might be used to make different anticipating instruments. K-implies is utilized to partition the information into groups in light of a formerly settled climate class. The conjecture looks at West Bengal's air contamination. The objective of this study is to give a weather conditions conjecture system to restrict the impacts of air contamination and to send off centered displaying computations for climate occasion expectation and determining. The creators directed many tests to assess the proposed approach. [9] utilized information mining strategies to gather climate information and uncover stowed away examples inside a tremendous dataset. To distinguish and foresee

a meteorological state, time series, K-mean grouping, and Nave Estimate were joined. This information mining approach is utilized to extricate information from a climate dataset in Gaza City. This information can be used to go with significant expectations for choice making, yet to advance powerfully, dynamic information mining techniques should be fabricated. Fast changes in climate and startling occasions can be predicted utilizing a powerful model.

2.4 RELATED WORKS

Various meteorologists and cautioning focuses gave themselves to this concentrate throughout the past hundred years, making progresses in observational innovation, heightening physical science, air climate communications, the environmental limit layer and air-ocean interface, sea reactions, and estimating strategies [11]. In any case, huge issues with prescient capacities persevere, especially with TC beginning, power, and hazard determining. As a rule, the most unmistakable hurricane dynamical estimate models have low precision, inferable from incorrect vortex introduction of TCs, deficient depiction of muddled actual cycles, and coarse goal [12,13].

As per some exploration, erroneous portrayals of the air-sea energy stream under very high wind speed conditions would make it challenging to imitate the power of TCs effectively [14]. Moreover, there is a wide agreement that upper sea input has huge effect on TCs, albeit hardly any functional mathematical estimate models consider it, diminishing model execution [15,16]. Moreover, different strategies, as measurable models, can't of managing the confounded and nonlinear connection between TC-related factors; subsequently, their conjecture results should be upgraded further [17–18].

Researchers have as of late explored the utilization of AI (ML) to examine satellite, radar, in-situ information, and different sources to further develop TC anticipating abilities. In light of their purposes, AI calculations can be arranged as man-made brainpower (man-made intelligence) in three ways: highlight choice, gathering, and relapse or grouping [19]. Feature selection algorithms can use attribute selection to exclude extraneous attributes from tasks, increasing task effectiveness and model correctness. A typical Exhaust disintegration strategy, for instance, can handle spatio-fleeting issues that the exemplary tensor deterioration calculation can't [20]. Grouping calculation, which can consequently segment an example dataset into numerous classifications, is one of the early methodologies utilized in design acknowledgment and information mining. This has various purposes in enormous information examination. The limited blended model (FMM) [21], progressive grouping [22], and K-implies technique are instances of normal bunching calculations [23].

Support vector machine (SVM) for characterization [24] and support vector backslide (SVR) for backslide [25] are two praiseworthy computations that may really deal with nonlinear issues by setting down segment abilities. Besides, choice tree (DT) [18] is a notable methodology for mining and showing gathering rules with high accuracy. By far most of arranging related work is finished with reproduced mind associations (ANNs), which are viewed as broad approximators for complex nonlinear mappings [26]. In 2006, Hinton, a notable man-made intelligence ace, proposed the significant mind network model, introducing another period of profound learning [28] and repetitive brain organizations (RNNs) [29] are instances of exemplary organizations. Profound learning (DL) enjoys upper hands over run of the mill AI calculations in high-layered information, making it more reasonable for troublesome applications. Subsequently, choosing the reasonable AI calculation for assorted information and requests could assist with resolving genuine issues all the more proficiently.

Given the advantages of AI in managing huge scope information, a few scientists tried different things with remote detecting, meteorology, and sea information. Remote detecting applications include: (1) arranging hyperspectral pictures and recognizing oddities utilizing profound brain organizations, for example, meager autoencoder (SAE), CNN, and RNN. (2) Utilizing engineered opening radar (SAR) pictures or high-goal satellite pictures to perform scene characterization, object location, picture recovery, and boundary reversal. (3) Making recovery strategies and forward models that are quick and precise. (4) Doing multi-source information combination and 3D reproduction. Honing the super-goal, combinations of capabilities and choice sets, and combinations of heterogeneous information are the fundamental obligations of multi-source information combination. Programmed tie point acknowledgment and matching are instances of 3D recreation challenges [30,31] [32] and direction entryway repetitive unit (TrajGRU) [33] are delegate for precipitation expectation, as is dissecting radar pictures in light of conventional brain organizations to anticipate transient future precipitation [34], as well as the arising utilization of profound brain networks for anticipating the air's dissipation pipe level at the sea surface [35]. AI in the marine field centers around recognizing mesoscale vortices or diminishing the dimensionality of satellite sea information. Many explorations have shown that profound learning organizations, for example, deepeddy [36], eddynet [37], and oednet [38], are very helpful for perceiving sea whirlpools.

AI applications in TC expectations can be parted into five classifications, as demonstrated in Figure 2.1. As far as TC beginning estimates, definitive objectives are to make constant probabilistic gauges of a particular locale and quantitative conjectures of the timing and area of twister beginning to screen the tropical sea more readily. Notwithstanding, as of now, AI can conjecture whether the forerunners will change into TCs, as well as the occasional recurrence of TC beginning in every area, which compares to a characterization challenge and a relapse task in AI, separately. For TC beginning expectation, scientists for the most part utilize DT, strategic relapse (LR), SVM, and outfit calculations like AdaBoost and arbitrary woodland (RF). Albeit these gathering calculations are possibly better than a solitary technique, they should in any case be assessed dependent upon the situation. Profound learning calculations, as multi-facet perceptrons (MLPs) and CNNs, that can all the more likely fit complex capabilities and decipher visual information, are likewise significant in further developing TC beginning foreseeing apparatuses.

AI based models are frequently worked from factual learning approaches for TC track conjectures. MLP and RNN, the analysts' favored techniques, have additionally been demonstrated to be more viable than more seasoned strategies. Moreover, in light of the fact that TCs change in both general setting, spatiotemporal models, for example, ConvLSTM have performed well in this expectation challenge. Beside those exemplary thoughts, there is likewise the endeavor to utilize generative antagonistic organizations (GANs) to create anticipated TC cloud pictures and afterward find the TC community on them, or the utilization of profound conviction organizations (DBNs) and bunching calculations, for example, FMM to look for generally comparative ways for guaging. Every one of these strategies delivered great trial brings about their distributions, however it is not yet clear whether they can create preferred functional foreseeing results over existing frameworks. DT, generally speaking based calculation, can be utilized to mine prescient elements and rules for TC landfall and curvature, giving the preparation to future track conjecture displaying.

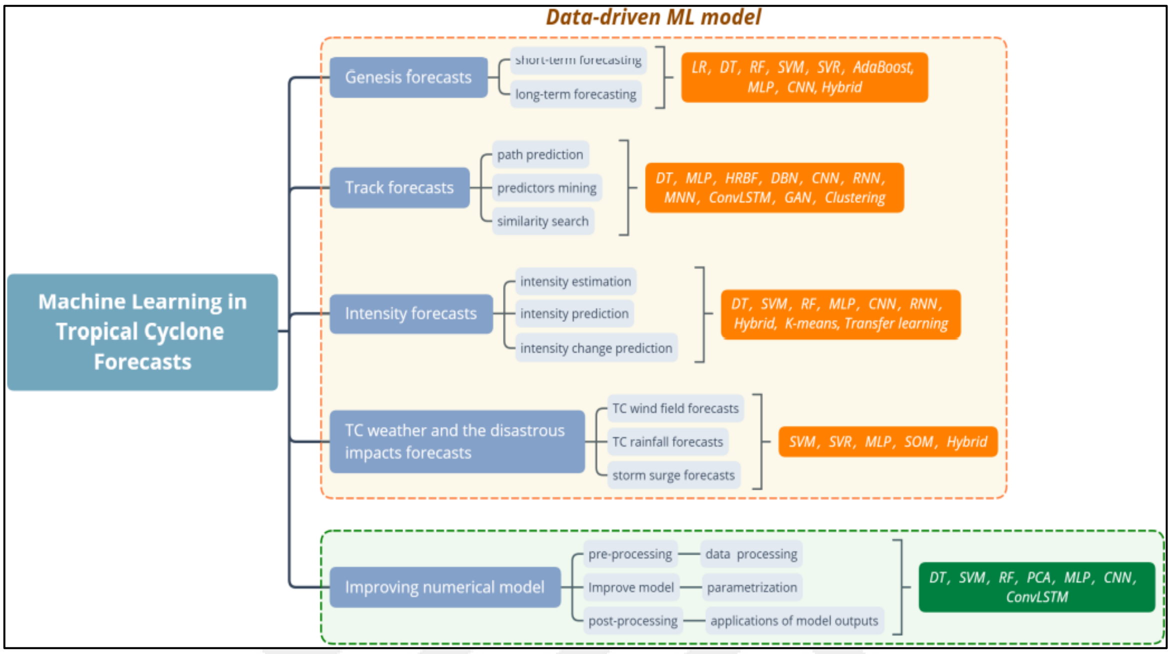

Figure 2.1: An Overview of Situations Employing Machine Learning in Tropical Cyclone Forecasting [38].

The outcomes showed that AdaBoost was the best strategy for typhoon beginning, with an estimate exactness of 97.2% (F1-score) contrasted with customary straight factual models utilizing natural variables connected with MCSs/TCs more than a 6 h expectation window. At the point when the lead time was expanded to 12, 24, or even 48 hours, heartiness was likewise guaranteed. Likewise, [43] utilized RF to make 2-hour estimates of MCS development and found that RF had serious areas of strength for a to identify MSCs. almost 100% exactness was accomplished for the 550 MSCs noticed (inside 50 km). Moreover, the DT calculation in [44] was applied to figure future hurricane arrangement utilizing tropical irritations portrayed by vorticity fields from the Naval force Worldwide Climatic Expectation Framework (NOGAPS). The program produced six order rules in view of natural elements, and the test informational index's expectation precision was 84.6%. Satellite information can be utilized as an information hotspot for indicators notwithstanding re-examination or reproduction information. For instance, [46] confirmed eight indicators from WinSAT satellite-estimated sea surface breeze and precipitation.

2.5 CHAPTER SUMMARY

Depending on the comprehensive literature review carried out in this work, it can be inferred that several research efforts, development, and engineering principles have been dedicated and implemented to enhancing and modifying the navigation system of robots under various operating conditions, which include indoor environment, outdoor surroundings, and places which are featured with unknown circumstances and large uncertainties. Several scholars and researchers designed mobile robots with enhanced navigation characteristics. They employed numerous intelligent algorithms and ML techniques to improve the path planning process and help robots avoid collisions with physical obstacles. In addition, using ML algorithms contributed to the significant enhancement and remarkable improvements in the accuracy and performance of the robot's navigation system. Consequently, this work investigates the use of some ML algorithms to improve the navigation system in mobile robots depending on numerical analysis and simulations.

3. PROPOSED SYSTEM

Weather classification and prediction is an important application of machine learning in the field of meteorology. Accurate weather prediction can help people prepare for extreme weather events, such as hurricanes, tornadoes, and blizzards. In recent years, various machine learning algorithms have been used to classify and predict weather conditions, including Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. Direct Relapse is a clear methodology for displaying the association between meteorological elements and the objective variable. KNN is a non-parametric strategy that can be utilized to distinguish weather patterns in view of likenesses among present and past meteorological information. The Choice Tree calculation is a notable procedure that can be utilized for both climate grouping and expectation. CNN is a profound learning technique that utilizations picture information to characterize weather patterns. The LSTM is a kind of intermittent brain network that can foresee weather conditions over the long haul. In this section, we will take a gander at how every one of these calculations can be utilized to characterize and foresee climate. We will describe the data preprocessing steps, the algorithmic implementation, and the performance evaluation metrics used for each algorithm. We will also compare the performance of each algorithm and discuss their respective strengths and weaknesses. Finally, we will discuss the potential applications of weather classification and prediction in various fields, including agriculture, transportation, and emergency management.

3.1 METHEDOLOGY

Weather forecasting has become increasingly essential as a result of its uses in a variety of industries, including agriculture, energy companies, and everyday life. Weather prediction has become a real-time challenge for the world in the previous decade. Because of the ever-changing meteorological conditions, forecasting is becoming more difficult. Weather forecasting is the prediction of how the current state of the atmosphere will change in the future. Understanding the numerous contributing elements that produce weather variations is crucial for efficient weather analysis. The process of recording meteorological data such as wind direction, wind speed, humidity, rainfall, temperature, and so on is known as weather forecasting. Since machine learning techniques are more robust to perturbations, in this project we applied linear regression and LSTM to predict the weather such as temperature, rainfall etc. and compare both approaches and analyzed it. We used two different datasets for the same. Coming to result that we got from each approaches was quite amazing. In the linear regression approach, we got mean absolute error about 96.32 mm and 2.69 celsius when performing rainfall and temperature prediction respectively whereas in the deep learning approach, the mean absolute error was 0.002268 degree celsius, 0.003266 km/h, and 0.003069 Pascal when performing temperature, wind speed and pressure prediction respectively. We could clearly see the difference between the outcomes.

In this section, we will describe the methodology for weather classification and prediction using Linear Regression, KNN, Decision Tree, CNN, and LSTM. The methodology consists of the Following :

a. Data Preprocessing: The first step in the methodology is to preprocess the weather data. This includes cleaning the data, removing any missing values, and normalizing the data. The data may also need to be transformed into a format that can be used by the specific algorithm being used.
b. Algorithmic Implementation: The next step is to implement the algorithm(s) being used for weather classification and prediction. This involves training the algorithm(s) on the preprocessed data, tuning the hyperparameters, and evaluating the performance of the algorithm(s).
c. Performance Evaluation: The final step is to evaluate the performance of the algorithm(s). This involves comparing the predicted weather conditions with the actual weather conditions, using metrics such as accuracy, precision, recall, and F1 score. For the weather forecasting, we try to solve the problem using two methods and those were:Machine Learning, Deep Learning
We divide the problem using two methods and we have done so because of the following reasons: -
a. Tackle the problem with different approaches.

b. Will know what the strengths and weaknesses of each method are.

c. Will know what the difficulties are one could face while approaching the problem with these two methods.

d. Will know what the efficiency of each process is.

Coming to the first approach that is Machine Learning, here we have solved the problem of weather forecasting with linear regression. In this approach, weather forecasting like rainfall prediction and temperature prediction is done separately. Also, the linear model which was trained for temperature predictions and rainfall predictions were trained using ordinary least square and regularization respectively.

Now coming to the second approach that is Deep learning, here we have solved the problem of weather forecasting in a single stroke. Mean, only in one solution, we have done the forecasting of the temperature, wind speed, and pressure. We have trained our model on LSTM (advanced version of the RNN).

3.1.1 Data Collection

There are different sites accessible now that permit clients to get climate perceptions or information from a great many geological areas (weather conditions stations, satellites, radars, and so on.). Climate APIs are Application Programming Points of interaction that permit anybody to interface with enormous climate data sets and get current and verifiable information. Numerous versatile applications have been created utilizing APIs and cellphones to convey hourly conjectures, extreme weather conditions cautions, and appropriate climate data for any area in the world. Different APIs were tried to guarantee they met the necessities in general.

Table 3.1: Dataset Description.

File	Rows	Col.
weather_data_train.csv	3140	16
weather_data_test.csv	3140	16
weather_data_train_labels.csv	1346	2
weather_data_test_labels.csv	1346	2

The files containing the training and testing data (first two files in above table) consisted of all the different parameters which affects the relative humidity and decides the classification of weather. The rest of the two files consisted of the labels which were either 0 for “not dry” and 1 for “dry”, and relative humidity associated for each row entry in the training data.

Initial analysis of training data was done by extracting features such as mean, standard deviation, minimum, maximum, etc.

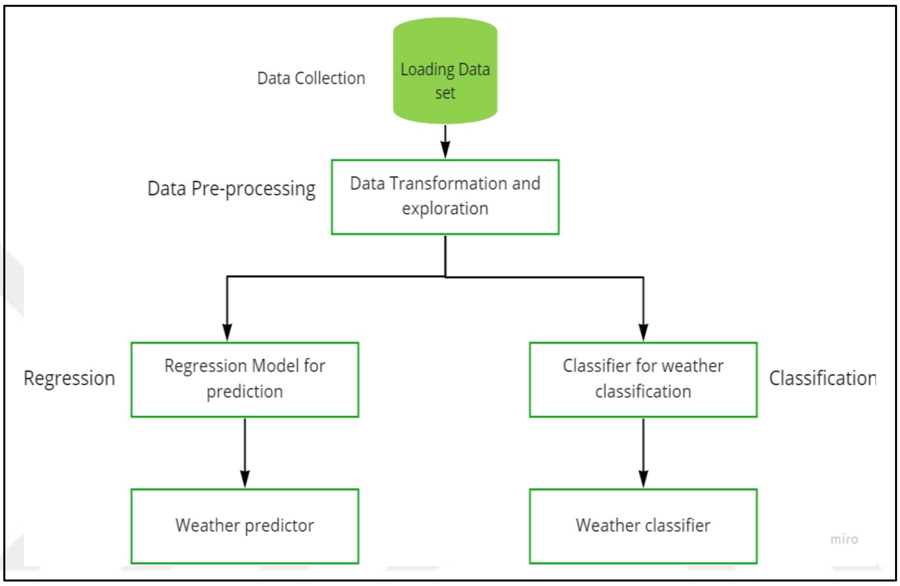

Figure 3.1: Methodology Block Diagram [40].

3.1.1.1 Data preprocessing

The initial stage of data mining is data collection and preprocessing. The crucial stage is data preprocessing because valid data will only yield accurate output:

We start preprocessing by standardizing the data using below given formula. We should standardize the variables before applying PCA because if one component (eg T_mu) If one feature (e.g. W) varies less than another due to their separate scales, PCA may conclude that the direction of maximal variance more closely aligns with the 'W' axis if those characteristics are not scaled. [7]

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

After standardizing the resultant data has a mean of zero and standard deviation of one. After this, we will be projecting the 16-dimensional weather data to three-dimensional principal components.

3.1.1.2 Feature extraction

Feature extraction is an important step in weather classification and prediction using Linear Regression, KNN, Decision Tree, CNN, and LSTM. The goal of feature extraction is to identify the most important features in the weather data that can be used to accurately predict the weather conditions. The specific feature extraction methods used for each algorithm are described below:

Linear Regression: For Linear Regression, the most important features are identified using the OLS method. The OLS method estimates the coefficients of the linear regression model by minimizing the sum of the squared residuals between the predicted and actual weather conditions. The coefficients with the highest absolute values are considered to be the most important features.

K-Nearest Neighbors (KNN): For KNN, the most important features are identified using feature selection methods such as chi-square, mutual information, and correlation-based feature selection. These methods calculate the relevance of each feature to the target variable, and select the features with the highest relevance.

Decision Tree: For Decision Tree, the most important features are identified using feature importance methods such as Gini impurity and entropy. These methods measure the purity of a node in the decision tree based on the distribution of the target variable, and select the features that result in the greatest decrease in impurity.

Convolutional Neural Networks (CNN): For CNN, the most important features are identified using convolutional layers. The convolutional layers perform feature extraction by convolving filters over the input data, and identifying the most important features based on the filters that produce the highest activation values.

Long Short-Term Memory (LSTM): For LSTM, the most important features are identified using time-series analysis methods such as autocorrelation and cross-correlation. These methods identify the temporal dependencies between the weather variables and select the variables with the highest dependencies. Additionally, feature selection methods such as mutual information and correlation-based feature selection can also be used to identify the most important features.

3.1.2 Regression

Regression analysis is a strong statistical technique for examining the relationship between two or more variables of interest. The regression procedure allows you to safely establish which elements are most important, which factors may be ignored, and how these factors influence each other. [11].

3.1.2.1 Linear regression

Linear regression is a statistical method that can be used for weather prediction. The basic idea behind linear regression is to fit a linear equation to the weather data, so that the equation can be used to make predictions about future weather conditions. The equation takes the form of: y = b0 + b1x1 + b2x2 + ... + bnxn.[9]

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \tag{3.2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \tag{3.3}$$

where y is the target variable (e.g. temperature, humidity), x1, x2, ..., xn are the input variables (e.g. time of day, day of the week, season), and b0, b1, b2, ..., bn are the coefficients that represent the relationship between the input variables and the target variable. To use linear regression for weather prediction, we first need to collect weather data for a specific location over a period of time. Temperature, mugginess, pressure, wind speed, and precipitation, as well as time and date data, can be in every way remembered for this information. The information can then be partitioned into two sets: a preparation set and a test set, with the preparation set being utilized to fit the direct relapse model and the test set being utilized to assess its exhibition. Once the model is fit, we can use it to predict future weather conditions by plugging in the input variables for the desired time and location. For example, we could predict the temperature at 3pm on a certain date based on the historical temperature, humidity, pressure, and time of day data. Straight relapse expects a direct connection between the info factors and the objective variable. Other relapse draws near, like polynomial relapse or non-straight relapse, might be more proper assuming that the relationship is non-direct. Additionally, linear regression may not be the best method for predicting extreme weather events or long-term weather patterns, as these are often influenced by complex atmospheric processes that cannot be accurately modeled using a simple linear equation.

3.1.2.2 Decision tree

Decision trees are another statistical method that can be used for weather prediction. Decision trees are a type of supervised learning algorithm that can be used for both classification and regression problems. In the context of weather prediction, decision trees can be used to predict the value of a continuous variable (e.g. temperature, humidity) based on a set of input variables (e.g. time of day, day of the week, season). Every penultimate hub conveys the last characteristic or various qualities for settling on an official choice of the issue in figure 2.

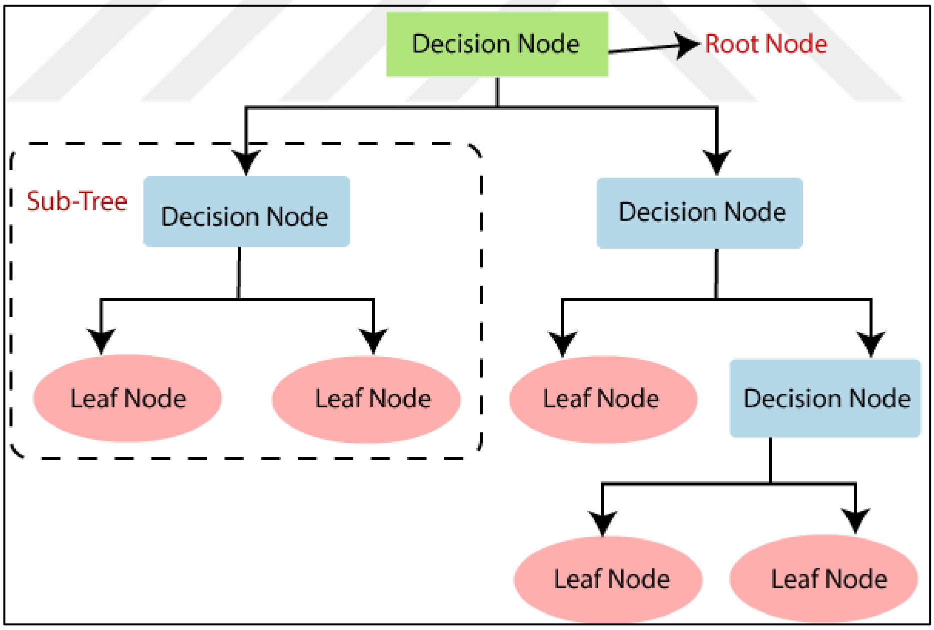

Figure 3.2: Decision Tree [44].

The basic idea behind decision trees is to partition the data into smaller and smaller groups based on the input variables, until the groups are as homogeneous as possible in terms of the target variable. The result is a tree-like structure that consists of a series of binary splits that divide the data into smaller and smaller groups. To use decision trees for weather prediction, we first need to collect weather data for a specific location over a period of time, as well as time and date information. We can then split the data into a training set and a test set, with the training set used to build the decision tree and the test set used to evaluate its performance. The decision tree is built using an algorithm that selects the input variable that best separates the data into two groups based on the target variable. This process is repeated recursively for each resulting subgroup until the groups are as homogeneous as possible in terms of the target variable, or until a stopping criterion is met (e.g. a maximum depth is reached). Once the decision tree is built, we can use it to predict future weather conditions by traversing the tree based on the input variables for the desired time and location. The last expectation is the normal or larger part vote of the objective variable qualities for the subgroup with the information factors. It is pivotal to feature that choice trees are defenseless to overfitting, which happens when the tree is excessively confounded and fits the clamor in the preparation information as opposed to the hidden connection between the info factors and the objective variable. To defeat this issue, approaches like pruning, regularization, and troupe strategies can be applied to further develop the choice tree's presentation.

3.1.3 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.[10] In this project, we aim to create a classifier which classifies weather into 0(not dry) or 1(dry). This kind of classification with two categories is called binary classification.

3.1.3.1 Artificial neural networks (ANN)

ANNs are a form of machine learning algorithm that is inspired by the structure and function of the human brain. ANNs are made up of interconnected nodes (also known as neurons) grouped into layers. Each neuron in a layer is connected to neurons in the previous layer and the next layer, with each connection having a weight that determines the strength of the signal transmitted between the neurons. ANNs can be used for a wide range of applications, including image recognition, natural language processing, and time series prediction, such as weather forecasting. In the context of weather forecasting, Based on previous data, ANNs may forecast variables such as temperature, humidity, and wind speed. To use ANNs for weather prediction, we must first gather weather data for a specific place over time.

Temperature, humidity, pressure, wind speed, and precipitation, as well as time and date information, can all be included in this data. The data can then be divided into two sets: a training set and a test set, with the training set being used to train the ANN and the test set being used to evaluate its performance.

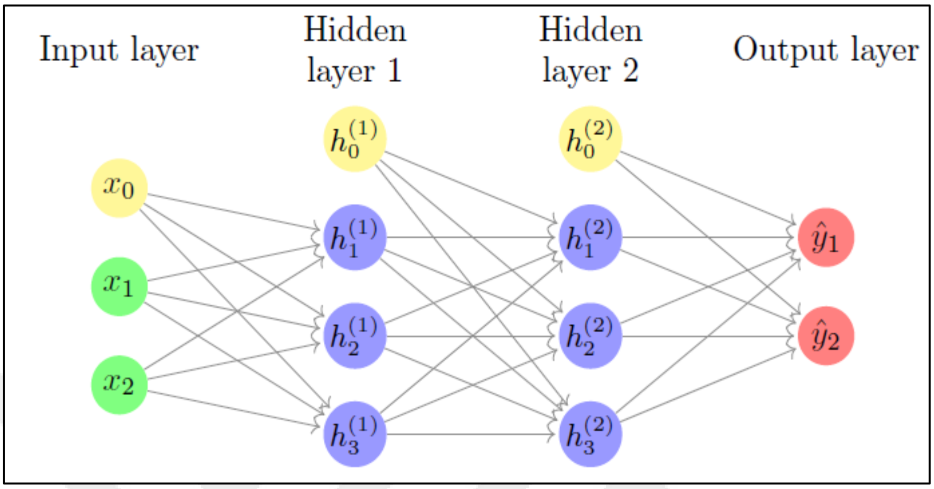

Figure 3.3: Neural Network Architecture Diagram [47].

The preparation of ANNs involves changing the loads of the associations between neurons to decrease the inconsistency among expected and genuine qualities. This is in many cases achieved using an advancement calculation like slope plunge. To build the model's presentation, the ANN configuration, including the quantity of layers and neurons, as well as the enactment capability used for every neuron, can be adjusted. Once the ANN is trained, we can use it to predict future weather conditions by inputting the historical data for the desired time and location. The output of the ANN will be the predicted values of the target variables for the given input. One advantage of ANNs over traditional statistical methods is their ability to capture complex nonlinear relationships between the input variables and the target variable. However, ANNs can also be prone to overfitting, particularly if the model is too complex or if there is not enough training data. Techniques such as regularization, early stopping, and dropout can be used to address this issue. Additionally, the interpretability of ANNs can be a challenge, as it is often difficult to understand how the model arrived at its predictions.

3.1.3.2 Lstm

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is designed to handle the vanishing gradient problem in traditional RNNs. At the point when the angle (i.e., the subsidiary of the blunder concerning the loads) turns out to be tiny during backpropagation, it becomes challenging to actually refresh the loads and train the model. Memory cells are utilized in LSTM organizations to store data after some time, while doors are utilized to oversee the progression of data into and out of the memory cells. The info door, neglect entryway, and result entryway are the three sorts of entryways in a LSTM organization. The information door directs the progression of information into the memory cell, the neglect entryway manages the progression of information out of the memory cell, and the result door controls the progression of information from the memory cell to the result. LSTM networks are particularly useful for time series prediction tasks, such as weather forecasting, where the input data has a temporal dimension. In the context of weather forecasting, LSTM networks can be used to predict variables such as temperature, humidity, and precipitation based on historical data.

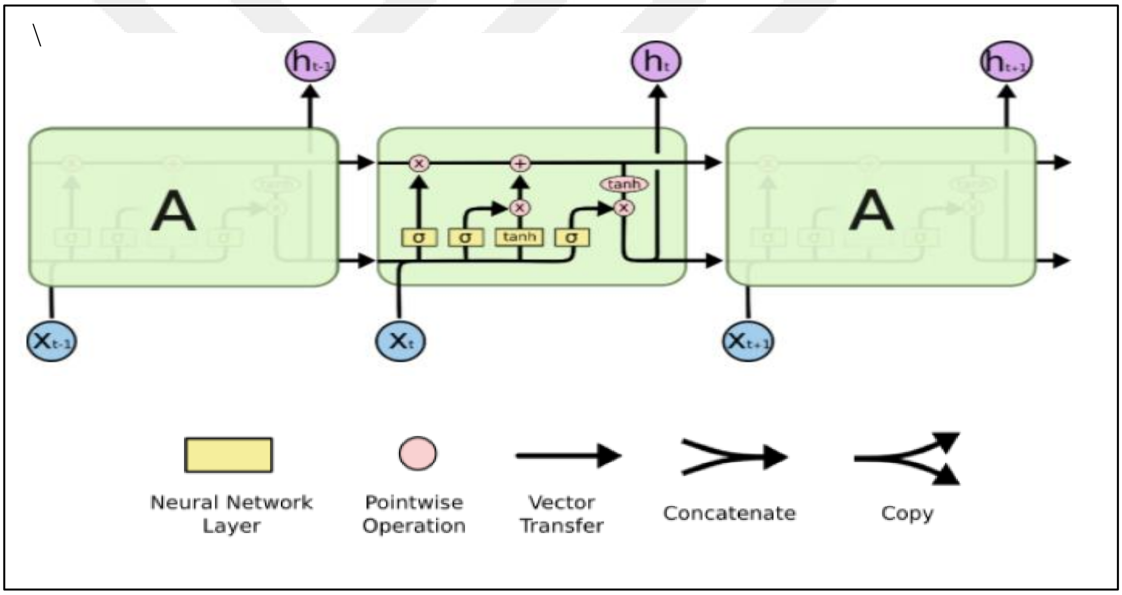

Figure 3.4: LSTM Architecture [48].

We have three different gates that regulate information flow in an LSTM cell. A forget gate, input gate, and output gate.

3.1.3.3 Forget gate

Long Short-Term Memory (LSTM) networks rely heavily on the forget gate. It is in charge of managing the flow of information from the memory cell in the previous time step to the memory cell in the current time step. There are three sorts of gates in an LSTM network: input gates, forget gates, and output gates. The forget gate decides which information from the previous time step to discard and which to keep. The forget gate takes the previous hidden state and the current input as input and applies a sigmoid function to the result. The sigmoid function returns a value between 0 and 1, indicating how much information to forget from the previous time step. If the output of the sigmoid function is 1, then all of the information from the previous time step is retained. If the output is 0, then all of the information is discarded. If the output is somewhere in between, then some of the information is kept and some is discarded. The forget gate allows LSTM networks to selectively remember or forget information from the previous time step, which is particularly useful for time series prediction tasks where the input data has a temporal dimension. By selectively remembering and forgetting information, LSTM networks can effectively capture long-term dependencies in the input.

3.1.3.4 Input gate

In (LSTM) organizations, the info entryway is a basic part. It is accountable for dealing with the progression of data in the ongoing time step from the ongoing contribution to the memory cell. There are three kinds of entryways in a LSTM organization: input doors, neglect doors, and result doors. The information entryway chooses which data from the ongoing contribution to keep and which to dismiss. The information entryway acknowledges the past secret state and the ongoing contribution as info and applies a sigmoid capability to the outcome. The sigmoid capability returns a worth somewhere in the range of 0 and 1, showing how much data to hold from the ongoing information. The input gate also applies a tanh function to the concatenation of the previous hidden state and the current input. The tanh function outputs a value between -1 and 1, which represents the new candidate values to be added to the memory cell [48]

3.1.3.5 Output gate

The output gate is a key component of Long Short-Term Memory (LSTM) networks. It is responsible for controlling the flow of information from the memory cell to the current hidden state, which is the output of the LSTM network.

There are three kinds of entryways in a LSTM organization: input doors, neglect doors, and result entryways. The result entryway chooses the data from the memory cell to yield in the ongoing time step. The result door takes the past secret state and the ongoing contribution as info and applies a sigmoid capability to the result. The sigmoid capability returns a worth somewhere in the range of 0 and 1, demonstrating the amount of information to be yield from the memory cell. The ongoing memory cell is likewise exposed to a tanh capability through the result door. The tanh capability returns a worth between - 1 and 1, showing the ongoing competitor yield. The result door then duplicates the sigmoid capability yield by the tanh capability result to decide the new secret state to be yield by the LSTM organization. The result door empowers LSTM organizations to yield important data from the memory cell at the ongoing time step in view of its pertinence to the current task. LSTM organizations can effectively catch long haul conditions in input information by choosing yielding pertinent data, which is basic for right expectations. The result entryway is just a single part of the LSTM organization, yet it is basic to the model's ability to learn and estimate precisely [49].

3.1.3.6 Cell state

We ought to now have an adequate number of information to ascertain the cell status. The cell state is then pointwise increased by the fail to remember vector. Whenever duplicated by values near nothing, this can possibly drop values in the cell state. The result of the information entryway is then used to play out a pointwise expansion, which refreshes the cell state to new qualities that the brain network sees as significant. This outcomes in our new cell state [50].

3.1.3.7 Conventional neural network

Convolutional Neural Networks (CNNs) have been widely used for various image classification tasks. However, they can also be used for weather prediction by treating weather data as an image.

In the context of weather prediction, the input data can be represented as an image with each pixel representing a feature such as temperature, humidity, pressure, or wind speed. The spatial correlation between these features can be captured by applying convolutional filters to the input image, which enables the CNN to learn patterns and features that are relevant for weather prediction. The CNN architecture for weather prediction typically consists of multiple convolutional layers followed by a few fully connected layers. The convolutional layers use filters to extract features from the input image and reduce the spatial dimensions of the image while preserving the important features. The fully connected layers are responsible for combining the extracted features and making the final predictions. To train the CNN, the input data is first preprocessed to ensure that all features are in a similar range of values. The data is then split into training and validation sets, and the CNN is trained using backpropagation and stochastic gradient descent optimization. Once the CNN is trained, it can be used to make weather predictions by feeding in new weather data as an image and obtaining the predicted output from the output layer of the network. The predictions can then be compared to the actual weather observations to evaluate the accuracy of the model. Overall, CNNs can be a powerful tool for weather prediction, especially when dealing with large datasets and complex patterns. However, they do require significant computational resources and training data, and their performance may be impacted by the quality and representativeness of the input data.

3.2 CONCLUSION

In this project we are predicting the relative humidity and classifying weather on parameters like temperature, pressure, air speed, etc. The most surprising, yet expected, finding in the project was the low correlation of variables with the overall weather. This study proposes and gives an efficient and exact climate expectation and estimating model in view of straight relapse ideas. We likewise utilize the ordinary condition model. These thoughts are essential for AI. The direct relapse condition is a superb weather conditions estimating model. We use factors like temperature, moistness, and wind speed. This can be utilized to make exact weather conditions estimates. This worldview likewise helps with regular direction. At the point when used to cleaner and bigger datasets, it can deliver unrivaled outcomes. Prior to handling of the datasets can be helpful in expectation, and natural information can likewise thwart the model's effectiveness.

4. RESULT AND DISCUSSION

4.1 EXPERIMENT SETUP

Since weather conditions estimating calls for time series, k-crease cross-approval is an incapable methodology for deciding whether our model will sum up to an autonomous test set. All things being equal, as displayed in table II, a 4-overlap forward fastening time-series cross approval was embraced, with the test set comprising of information from the year promptly following the preparation set. Since the model is based on past information and expectations on future information, this strategy all the more unequivocally recreates the climate at forecast time. An expectation to learn and adapt can likewise be built, which gives an important sign of the model's reliance on the size of the preparation set.

In light of this, the utilitarian relapse model boundaries were decided to limit the rms mistake in condition 9 arrived at the midpoint of over each of the four test sets in table II. We set the loads w2 = w3 = 1 in condition 3 since we thought about that distinctions in most extreme and least temperatures ought to be given equivalent weight. Since the utilitarian type of the assessor f in condition 7 was excessively huge for stochastic slope drop, a comprehensive network search was utilized to streamline the extra loads w1, w4, and w5. To augment every one of these boundaries, thorough lattice look through across the loads w and the quantity of neighbors k were done on the other hand. To get good guesses of these loads, an underlying comprehensive matrix search with huge augmentations was finished, with the upsides of each weight picked from the reach 0-50 in additions of 10. k = 5 neighbors were picked as the quantity of neighbors. This brought about starter assessments of w1 = 20 and w4 = w5 = 0. w4 = 0 was the ideal mean moistness weight, most likely in light of the fact that dampness related inadequately with greatest and least temperatures, and stickiness would be a more valuable indicator of precipitation. w5 = 0 was viewed as the best weight of the mean climatic strain since there were just minor deviations in the environmental tension that didn't seem, by all accounts, to be related with the greatest and least temperatures. With this in mind, the mean humidity and the mean atmospheric pressure were removed as features.

The hyperparameter of the quantity of neighbors k was then picked likewise, with a comprehensive network search across both steady and corresponding to informational collection size values. Factors of k in the reach 5-50 of every 5 augmentations were researched, as were upsides of k corresponding to the informational index size with proportionality steady in the reach 0.05-0.50 in 0.05 additions. Taking k proportionate to the size of the informational collection outflanked keeping k consistent, and the best proportionality steady was 0.10. The proportionality constants w1 and k were then tweaked at the same time with a last comprehensive framework search, with w1 taken from the reach 15-25 in augmentations of 1, and k taken from the reach 0.05-0.15 in augmentations of 0.05. This came about in w1 = 18 and k = 0.095|D|, where |D| is the quantity of data of interest.

4.2 DATA ANALYSIS

The files containing the training and testing data (first two files in above table) consisted of all the different parameters which affects the relative humidity and decides the classification of weather. The rest of the two files consisted of the labels which were either 0 for “not dry” and 1 for “dry”, and relative humidity associated for each row entry in the training data. Initial analysis of training data was done by extracting features such as mean, standard deviation, minimum, maximum, etc. as shown in Figure 4.1

	T_var	Po_var	P_var	Ff_var	Tn_var	Tx_var	VV_var	Td_var
count	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000
mean	4.466326	4.062051	4.058249	1.345066	3.662163	4.132288	98.008192	2.768010
std	5.254302	7.278152	7.274098	1.255878	7.034635	5.929607	106.710901	4.497696
min	0.016429	0.005000	0.005714	0.000000	0.000000	0.000000	0.000000	0.008095
25%	0.928571	0.492604	0.500000	0.571429	0.125000	0.382500	20.408036	0.582143
50%	2.568839	1.516964	1.519107	0.982143	0.980000	2.000000	55.928571	1.313125
75%	6.169955	4.395759	4.418750	1.696429	3.920000	5.780000	140.008884	3.089107
max	66.571250	108.117143	108.515536	17.571429	93.845000	105.125000	645.760000	77.449821

	T_mu	Po_mu	P_mu	Ff_mu	Tn_mu	Tx_mu	VV_mu	Td_mu
count	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000
mean	6.780096	758.805939	759.148643	3.674008	4.888025	8.661810	25.952053	3.373812
std	8.595021	8.521047	8.525485	1.222216	8.595641	8.910328	12.230215	8.083700
min	-19.312500	725.525000	725.875000	1.000000	-26.200000	-16.950000	0.625000	-22.912500
25%	0.987500	753.471875	753.821875	2.750000	-0.250000	2.250000	16.421875	-1.612500
50%	6.537500	758.850000	759.187500	3.500000	4.750000	8.125000	24.875000	3.118750
75%	13.953125	764.178125	764.525000	4.375000	11.950000	16.200000	35.375000	10.100000
max	25.787500	790.425000	790.812500	9.750000	23.150000	28.500000	50.000000	21.462500

Figure 4.1: Extracting Features.

A histogram is the most commonly used graph to show frequency distribution. Histogram can proveto be extremely useful when the data is numerical and we want to analyze the shape of data's distribution.

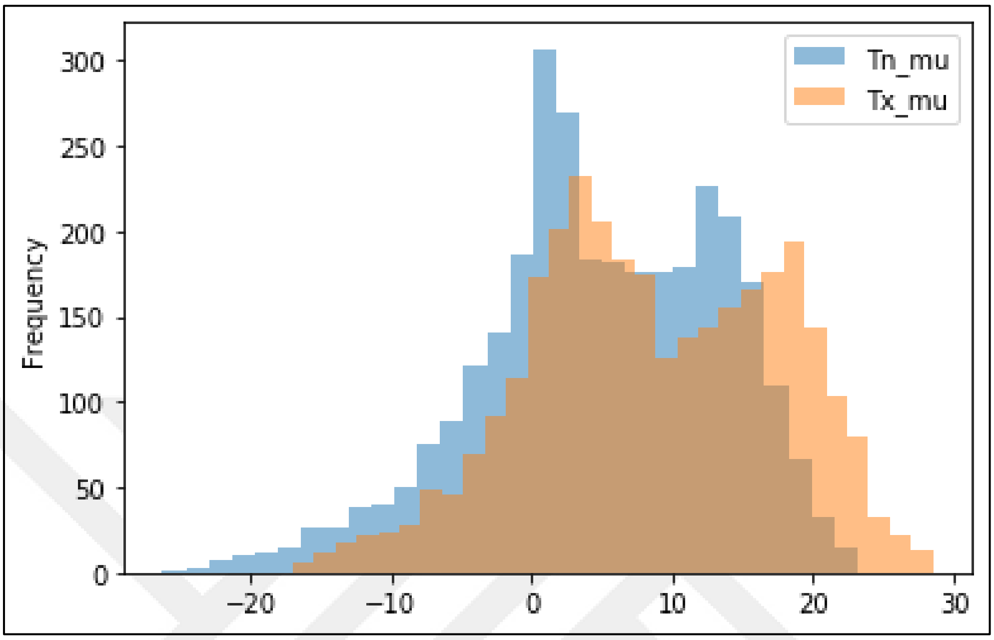

Figure 4.2: The Frequency of Data Distribution.

Above plot figure 4.2 is of the frequency distribution of Tn_mu and Tx_mu. We can see that the distributions have two maximums i.e Bimodal (double-peaked) Distribution which represents the two extreme seasons - summer and winter. Also, the distribution of both variables follow a similar trend. This is expected because both of them have a high correlation which means that if one rises/falls, the other rises/falls too. We can also see that the frequency of first maxima in both graphs is greater than frequency of second maximum which tells us that the number of cold days are more than hot ones.

A pair plot allows us to see both distribution of single variables and relations between two variables together. Pair plots are a great method to identify trends for follow-up analysis.

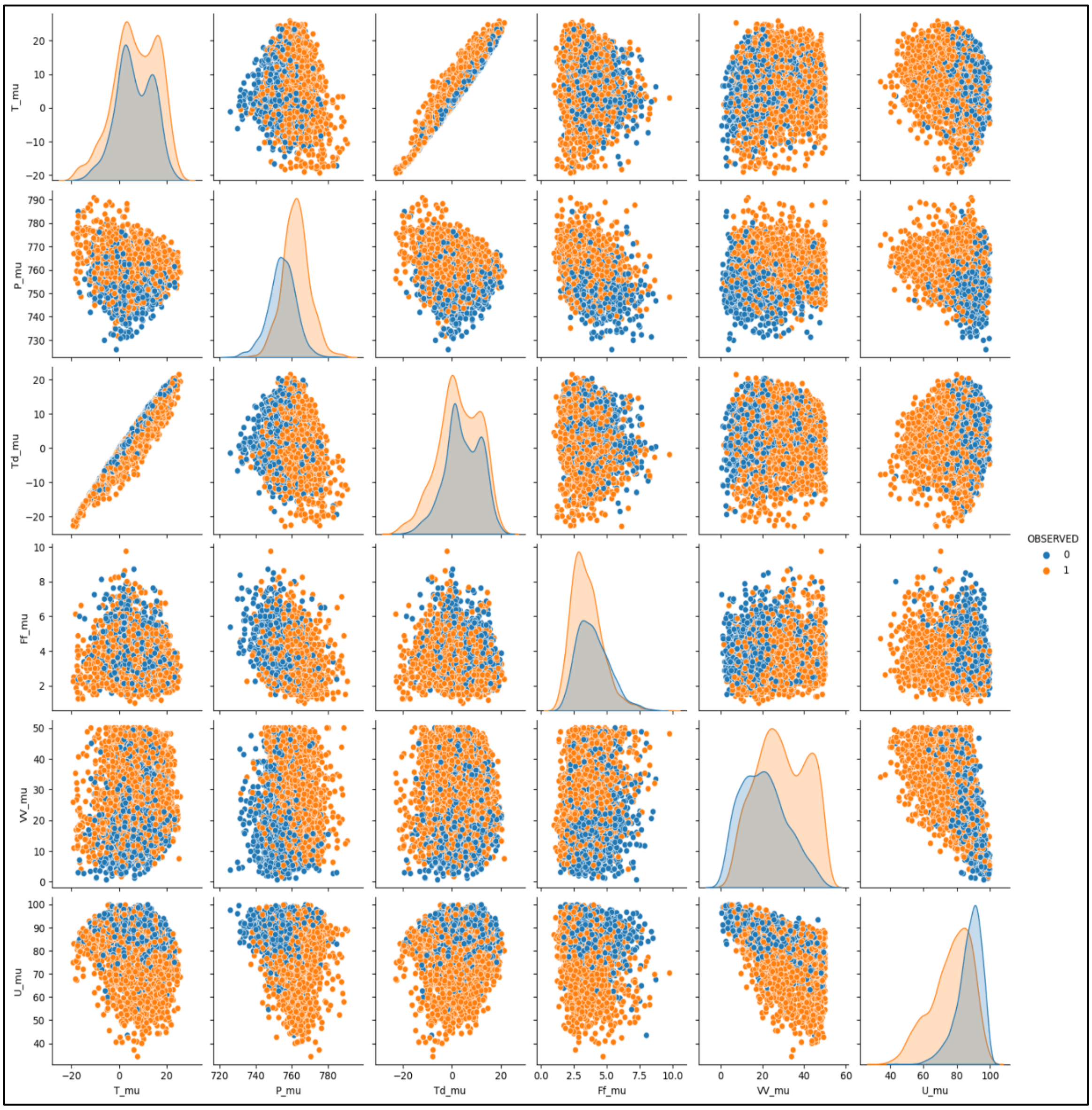

Figure 4.3: Pair Plots for T_mu, P_mu, Td_mu, Ff_mu, VV_mu, U_mu.

From the above pair plots figure 4.3 we can see that most of the data is randomly yet closely distributed without any significant correlation with each other. The only pair plot with any visible correlation is of T_mu and Td_mu which is expected because of their strong linear relationship in the real world.

The next step would be to see the correlation of different parameters with each other. The correlation tells us the relation between two variables. Positive correlation means that both the variables move in the same direction whereas negative correlation means that they move In opposite directions.[5] We calculate correlation coefficient (**r**) using formul :

$$\mathbf{r} = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}} \tag{3.4}$$

As we can see from the correlation matrix figure 4.4, some variables have low correlation with each other while some variables hold high correlation with each other. This high correlation between variables brings about a redundancy in the information that can be gathered by the data set. This redundancy can be removed by dimensionality reduction which is described in the next section.

	T_mu	Po_mu	P_mu	Ff_mu	Tn_mu	Tx_mu	VV_mu	Td_mu	T_var	Po_var	P_var	Ff_var	Tn_var	Tx_var	VV_var	Td_var	OBSERVED	U_mu
T_mu	1	-0.0634754	-0.0647653	-0.129089	0.99202	0.993854	0.197904	0.955727	0.144131	-0.195955	-0.196021	-0.0209119	0.211934	0.0566748	-0.0285134	-0.178432	0.0305501	-0.260787
Po_mu	-0.0634754	1	0.999989	-0.335656	-0.0899721	-0.0440898	0.0735027	-0.146007	0.225118	-0.283857	-0.283139	-0.240412	0.12575	0.134974	-0.243616	0.0231771	0.440013	-0.269538
P_mu	-0.0647653	0.999989	1	-0.335417	-0.0912599	-0.0453717	0.0731941	-0.147263	0.22503	-0.283367	-0.282725	-0.240176	0.125396	0.135071	-0.243537	0.0235145	0.439918	-0.269277
Ff_mu	-0.129089	-0.335656	-0.335417	1	-0.10194	-0.149504	0.10937	-0.14673	-0.24645	0.361483	0.361153	0.409312	-0.174823	-0.126245	0.11764	0.11022	-0.192452	-0.0571795
Tn_mu	0.99202	-0.0899721	-0.0912599	-0.10194	1	0.979744	0.167064	0.961942	0.0510356	-0.189527	-0.189645	-0.0295696	0.165914	-0.0108765	-0.0287533	-0.230666	0.00181042	-0.213275
Tx_mu	0.993854	-0.0440898	-0.0453717	-0.149504	0.979744	1	0.232505	0.934206	0.202114	-0.1979	-0.197961	-0.0146687	0.244814	0.0847651	-0.0313869	-0.148372	0.0566081	-0.311506
VV_mu	0.197904	0.0735027	0.0731941	0.10937	0.167064	0.232505	1	0.0217027	0.221391	0.0111468	0.0113873	0.0748892	0.149622	0.128403	0.10972	0.105754	0.340921	-0.640893
Td_mu	0.955727	-0.146007	-0.147263	-0.14673	0.961942	0.934206	0.0217027	1	0.00748227	-0.175056	-0.175116	-0.0427874	0.104634	-0.0136028	0.0223929	-0.254667	-0.098111	0.0319934
T_var	0.144131	0.225118	0.22503	-0.24645	0.0510356	0.202114	0.221391	0.00748227	1	-0.073988	-0.0734966	0.0533737	0.581422	0.648124	-0.0424512	0.465716	0.28378	-0.442629
Po_var	-0.195955	-0.283857	-0.283367	0.361483	-0.189527	-0.1979	0.0111468	-0.175056	-0.073988	1	0.999595	0.358113	-0.0463868	-0.0233537	0.190429	0.144366	-0.156247	0.0907483
P_var	-0.196021	-0.283139	-0.282725	0.361153	-0.189645	-0.197961	0.0113873	-0.175116	-0.0734966	0.999595	1	0.358572	-0.0456165	-0.0239272	0.190313	0.145447	-0.155696	0.0907656
Ff_var	-0.0209119	-0.240412	-0.240176	0.409312	-0.0295696	-0.0146687	0.0748892	-0.0427874	0.0533737	0.358113	0.358572	1	0.0380645	0.0277451	0.202818	0.234573	-0.11958	-0.0710154
Tn_var	0.211934	0.12575	0.125396	-0.174823	0.165914	0.244814	0.149622	0.104634	0.581422	-0.0463868	-0.0456165	0.0380645	1	0.12621	-0.0631088	0.153823	0.197789	-0.34158
click to hide r	0.0566748	0.134974	0.135071	-0.126245	-0.0108765	0.0847651	0.128403	-0.0136028	0.648124	-0.0233537	-0.0239272	0.0277451	0.12621	1	0.0080526	0.330586	0.166531	-0.232617
VV_var	-0.0285134	-0.243616	-0.243537	0.11764	-0.0287533	-0.0313869	0.10972	0.0223929	-0.0424512	0.190429	0.190313	0.202818	-0.0631088	0.0080526	1	0.134067	-0.296412	0.164174
Td_var	-0.178432	0.0231771	0.0235145	0.11022	-0.230666	-0.148372	0.105754	-0.254667	0.465716	0.144366	0.145447	0.234573	0.153823	0.330586	0.134067	1	0.0144774	-0.209823
OBSERVED	0.0305501	0.440013	0.439918	-0.192452	0.00181042	0.0566081	0.340921	-0.098111	0.28378	-0.156247	-0.155696	-0.11958	0.197789	0.166531	-0.296412	0.0144774	1	-0.44346
U_mu	-0.260787	-0.269538	-0.269277	-0.0571795	-0.213275	-0.311506	-0.640893	0.0319934	-0.442629	0.0907483	0.0907656	-0.0710154	-0.34158	-0.232617	0.164174	-0.209823	-0.44346	1

Figure 4.4: The Correlation Matrix.

4.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

The idea behind PCA is simply to find a low dimension set of axes that summarizes data. It is used when we need to tackle the curse of dimensionality among data with linear relationships. It helps to reduce the computational and cost complexities by transforming the original variables to the linear combination of these variables which are independent. [6]

PCA with 3 components is applied to reduce the dimensionality of the training data. We can see that the 3 components (Figure 4.5) explains around 61% of the original data. Can this value be higher for 3 components? Yes, if the correlation between the variables is higher.

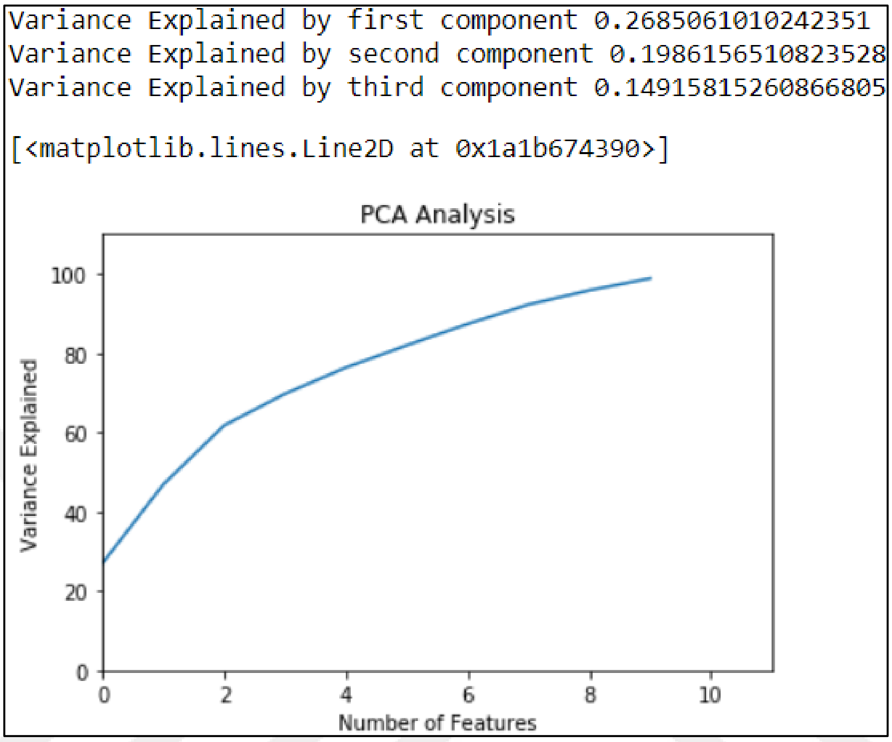

Figure 4.5: The 3 Components.

4.4 REGRESSION RESULTS

4.4.1 Linear Regression Results

We will now try to train a linear model with the provided training data. The train model is then tested on a test model and the analysis of error is presented below Figure 4. 6 and Table 4.1

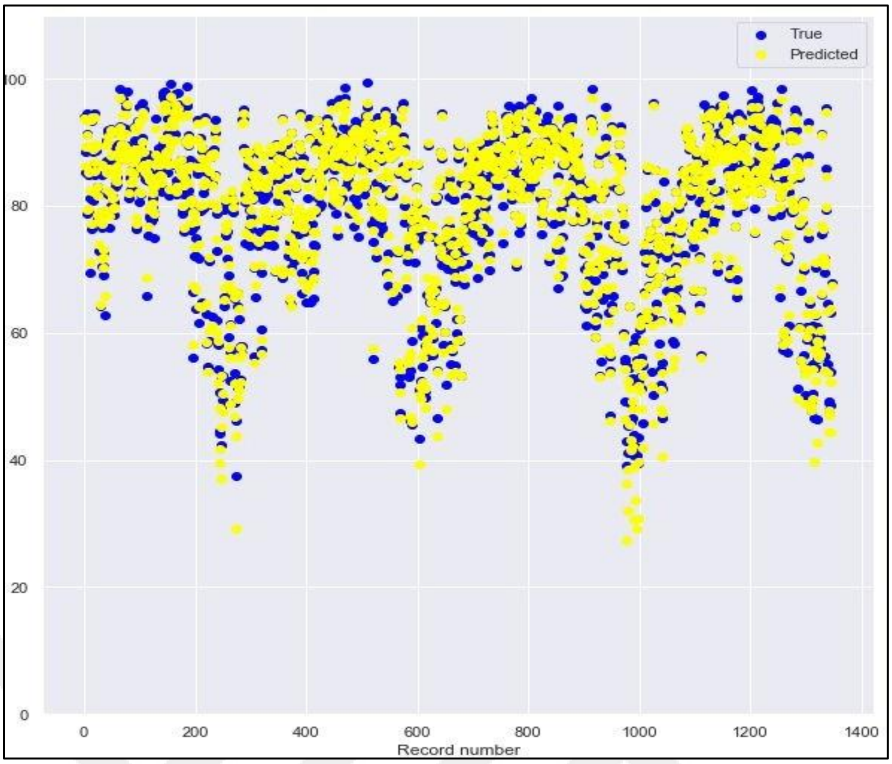

Figure 4.6: Linear Regression Results.

From the above table 4.1 we can see that the model performs well. But why? Well, the reason is given in the pair plot. If you see the pair plot closely, we can see the distribution of various parameters with relative humidity is randomly yet closely distributed without any particular direction. In such cases fitting a polynomial function through the data actually proves to work better because it reduces the overall error. Increasing the degree of the polynomial may seem to decrease the error on the training data but in reality it just overfits and thus work poorly on test data.

Table 4.1: Model Performance.

Result	
RMSE	1.3717
R-Square	0.99

We now try to see which of the 16 parameters affect the weather significantly. For this, we can look at the correlation matrix and try to rule out some parameters with low correlation with relative humidity. Why would this work? Well, correlation between two variables measures the strength of the linear relationship between them. Since we are training a linear model it makes sense to remove those variables with lesser correlation with relative humidity. From the correlation matrix we can see that variables VV_mu, T_mu, Td_mu have the highest correlation with relative humidity. Now we will train the regression model with these parameters and see the RMSE and R-Square values. After that we will marginally add other variables and repeat the analysis to see how much their addition contributes towards increasing the accuracy.

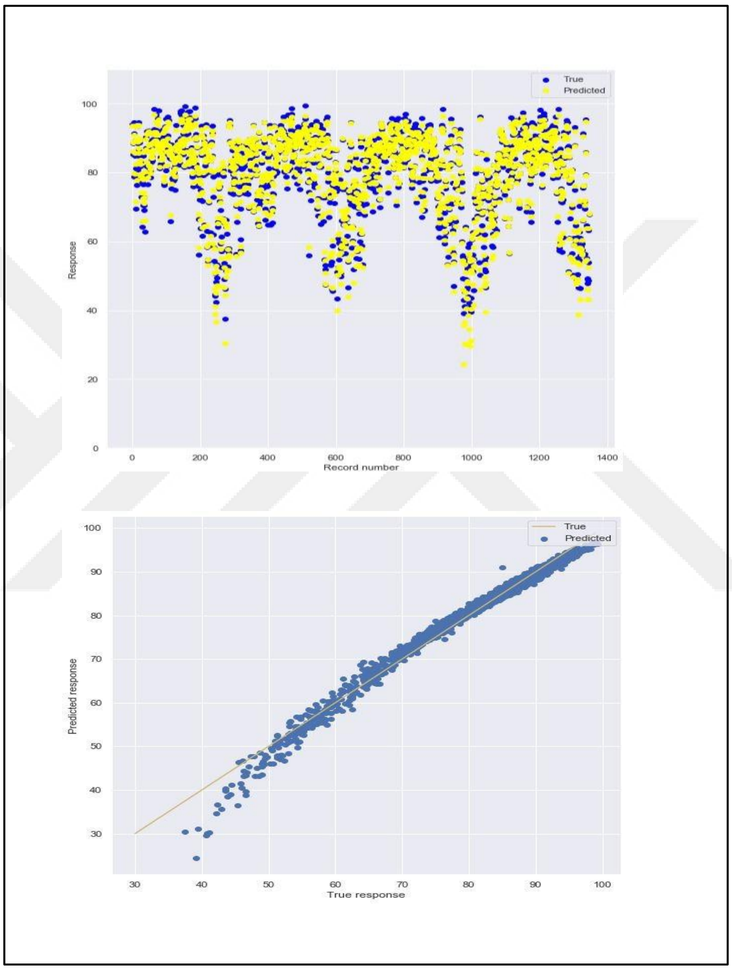

Figure 4.7: Linear Regression Results with PCA.

In the previous subsection we reduced the dimensionality by handpicking some of the features we thought affected the relative humidity significantly. Now we try to apply PCA and see how our model works. For this we train the model with different number of components and note the optimal number with low RMSE. As seen in given figure 4.7 there is a substantial decrease in RMSE when PCA equals 12. Therefore, optimal number of components for linear regression is 12.

The RMSE with 12 components is around 2, which is greater than what we got in linear regression. Why is that? It is mainly because when we apply PCA on data with low correlation, we lose most of the information.

4.4.2 Decision Tree

Decision trees are one of the popular algorithms used for building predictive models in machine learning. Weather prediction is one of the many applications of decision trees. Decision trees are used to classify weather conditions based on a set of input variables such as temperature, humidity, wind speed, etc. Here are the steps to build a decision tree for weather prediction:

Data Collection: Collect the data that contains information about the weather conditions of a particular region. The data should contain information about the temperature, humidity, wind speed, and other variables that may affect the weather.

Data Preprocessing: Clean the data by removing the missing values, outliers, and redundant variables.

Feature Selection: Identify the variables that are most relevant to predicting the weather conditions. Use techniques such as correlation analysis or feature importance ranking to determine the importance of each variable.

Splitting Data: Split the data into two sets, training and testing. The training set is used to train the decision tree, and the testing set is used to evaluate the performance of the model.

Building Decision Tree: Use the training data to build the decision tree using the decision tree algorithm. The decision tree algorithm selects the variable that is most relevant to the prediction and splits the data into two branches based on its value. This process is repeated until all the data is classified.

Evaluating the Model: Use the testing data to evaluate the performance of the decision tree model. Calculate the accuracy, precision, recall, and F1 score of the model.

Tuning the Model: Tune the decision tree model by adjusting the parameters such as the maximum depth of the tree, the minimum number of samples required to split a node, etc. to improve the performance of the model. Making Predictions: Use the trained decision tree model to predict the weather conditions of a new set of data.

In conclusion, decision trees can be used for weather prediction by building a model using the weather data and using it to predict future weather conditions. However, the accuracy of the model depends on the quality of the data and the effectiveness of the decision tree algorithm used.

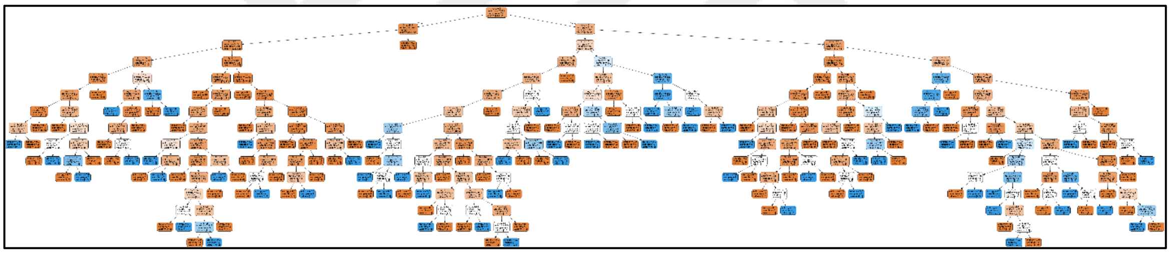

Figure 4.8: Decision Tree Model.

Table 4.2: Decision Tree Prediction Results.

date_time			
2013-07-10 08:00:00	34	34.0	0.0
2015-11-04 20:00:00	25	24.0	1.0
2015-09-21 09:00:00	34	34.0	0.0
2017-02-16 11:00:00	28	27.0	1.0
2012-07-21 01:00:00	28	28.0	0.0
...	...	...	...
2019-03-30 09:00:00	37	32.0	5.0
2015-11-12 12:00:00	32	32.0	0.0
2019-12-31 05:00:00	8	9.0	-1.0
2019-08-02 17:00:00	35	35.0	0.0
2019-10-22 08:00:00	26	26.0	0.0

4.5 CLASSIFICATION RESULTS

Weather classification is a common application of machine learning in the field of weather forecasting. In recent years, deep learning algorithms such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have shown promising results in weather classification. Here's how you can use CNN and LSTM networks for weather classification:

Data Collection: Collect the data that contains information about the weather conditions of a particular region. The data should contain information about the temperature, humidity, wind speed, and other variables that may affect the weather. In addition, label each data point with the corresponding weather condition, such as sunny, rainy, or cloudy.

Data Preprocessing: Clean the data by removing the missing values, outliers, and redundant variables. Then, split the data into two sets: training and testing.

CNN Model: CNN is a deep learning algorithm that is commonly used for image classification tasks. In the case of weather classification, you can represent the weather data as an image by plotting the different weather variables over time. Each image represents the weather conditions over a period of time. You can use a pre-trained CNN model such as VGG or ResNet to extract features from the images.

LSTM Model: LSTM is a sort of intermittent brain network that is reasonable for successive information like climate information. LSTM can catch the worldly conditions in the climate information, making it ideal for climate arrangement. You can utilize the separated elements from the CNN model as contribution to the LSTM model.

Combine CNN and LSTM: You can combine the CNN and LSTM models to create a hybrid model for weather classification. The CNN model can extract the spatial features from the weather data, and the LSTM model can capture the temporal dependencies.

Train the Model: Train the CNN-LSTM model using the training data. Use techniques such as data augmentation and regularization to prevent overfitting.

All in all, CNN and LSTM organizations can be utilized for climate order by addressing the climate information as pictures and utilizing a crossover CNN-LSTM model to separate spatial and worldly elements. The exhibition of the model relies upon the nature of the information, the adequacy of the pre-prepared CNN model utilized, and the design of the LSTM model.

4.5.1 Cyclone Detection by Convolutional Neural Network (CNN)

Cyclones are the large scale intense circular storm that originates over warm tropical oceans and is characterized by low atmospheric pressure, high winds and heavy rain. Variously referred to as hurricane or typhoon, cyclone has a significant socio-economic impact. The accurate forecast of the cyclone track has been a significant achievement of modern meteorology. Aided by the development of satellites and improved computational power, cyclone predictability has improved over the years. Recently with the advent of machine learning, attempts have been made to predict the cyclone genesis.

In this section I have tried to detect the presence of cyclone using the convolutional neural network. Relative vorticity and geopotential height at 500 hPa as well as mean sea level pressure and total cloud cover are used as inputs for the model. These fields are obtained from ERA_Interim reanalysis dataset, which is a 4X daily data with a resolution of 0.75 deg. In this study only the cyclones occuring in Atlantic basin is considered. A list of atlantic basin cyclonic disturbances can be found here (https://www.kaggle.com/noaa/hurricane-database) and (https://www.nhc.noaa.gov/data/). This data is used to create the target variable, which is assigned 1 if cyclonic disturbance is present and 0 otherwise. For simplicity I have considered all the disturbances regardless of its strength as cyclone. Data from the years 1979-2015 is used.

The following plot shows the mean sea level pressure (in color having the units of hPa) and relative vorticity at 500 hPa (only positive values are shown in red contours) during the various phases of cyclone development. All the plots are at 00 hrs UTC. The cyclone intensified around October 19th near Cayman Islands and progressed North-East crossing Florida hugging along the east coast of USA. It finally dissipated around October 27th.

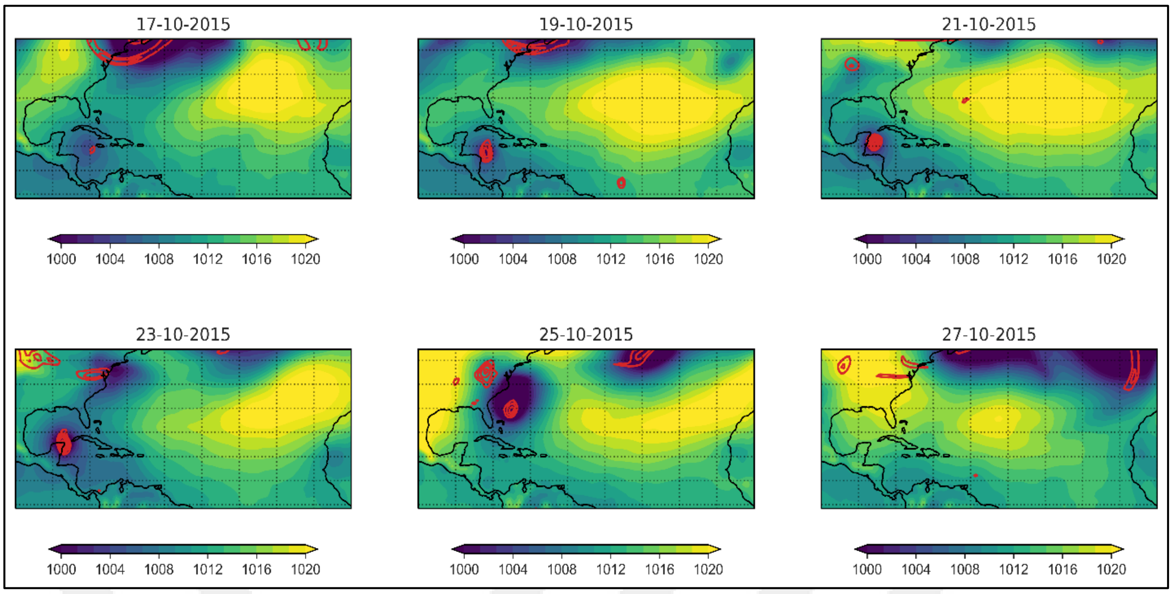

Figure 4.9: The Mean Sea Level Pressure.

Only about 20% of the times experience cyclonic activities. This uneven distribution of data could result in incorrect classification. For example, the model could simply classify all the outputs as non-cyclonic and still end up with 80% accuracy. So equal number of cyclonic and non-cyclonic times are considered. Training, validation and test data are split in the ratio of 60%, 20% and 20% respectively. Successive times will have a significant correlation. So different years are considered for training, validation and test split and is randomly shuffled among the individual subsets.

Three 2D convolution layers with max pooling is used. The output of convolution layers is flattened and fed to densely connected layers. The last layer has a sigmoid activation and classifies the output as either cyclonic or non-cyclonic event. Adam optimizer is used with binary crossentropy as the loss parameter. Model summary is printed below for reference.

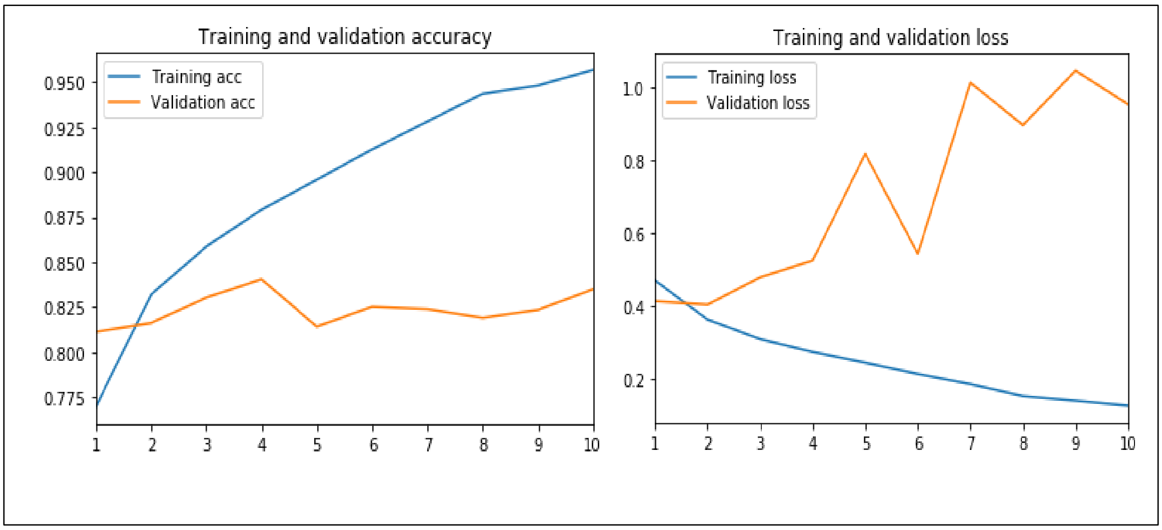

Figure 4.10: CNN Classification Results.

Validation accuracy is getting saturated around 84 %. I have only done an elemntary parameter tuning. Extensive tuning would probably lead to an even better performance. Now let us train the model with the enitre training and validation set. n_epochs is set to 4 as validation accuracy decreases after that.

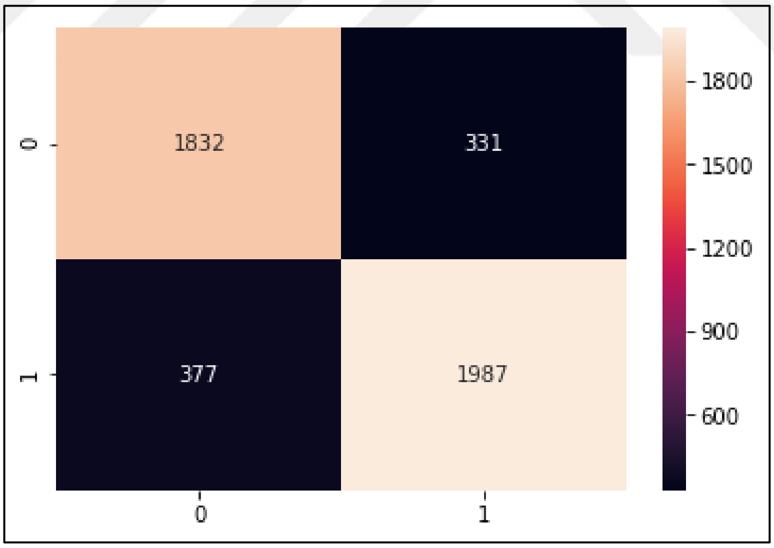

Figure 4.11: Confusion Matrix**.**

Test data has an accuracy of about 84 % similar to the validation set. Model seems to be perfoming fairly good. Further inclusion of additional parameters and better model tuning would perhaps lead to an even better performance.

4.5.2 Monsoon Classification by LSTM and Seasonal Arima Model

Indian Summer Monsoon Rainfall ISMR is the lifeline of Indian sub-continent. It accounts for about 70% of the India's annual rainfall. Agriculture of India is extremely dependent on monsoon and hence a failed monsoon would severely impact the country's economy. Thus a lot of research has been concentrated on the predictability of monsoon. One can track the state of the current monsoon season here.

In this section, an attempt has been made to predict the next month's mean precipitation using past year's data. An intermittent brain organization (RNN), explicitly Lengthy Transient Memory (LSTM) network is utilized. An other model which utilizes an Occasional AutoRegressive Coordinated Moving Normal (SARIMA) model is likewise utilized for examination.

SRIMA model is showing a seasonal periodicity of 12 months (as expected, the annual signal). More details can be read in the model summary below. Model diagnostics is plotted below. The residual looks like a gaussian white noise. Model seems to be fitting the data well.

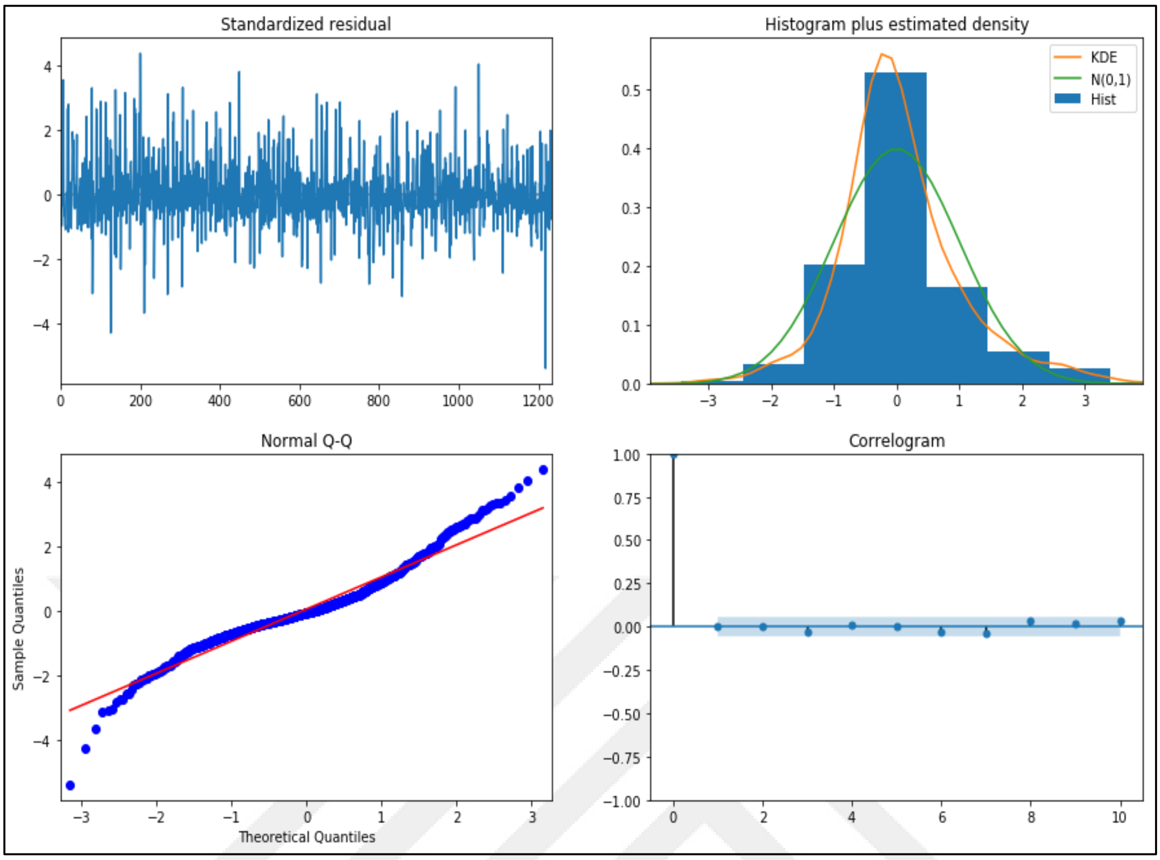

Figure 4.12: SRIMA Model Diagnostics.

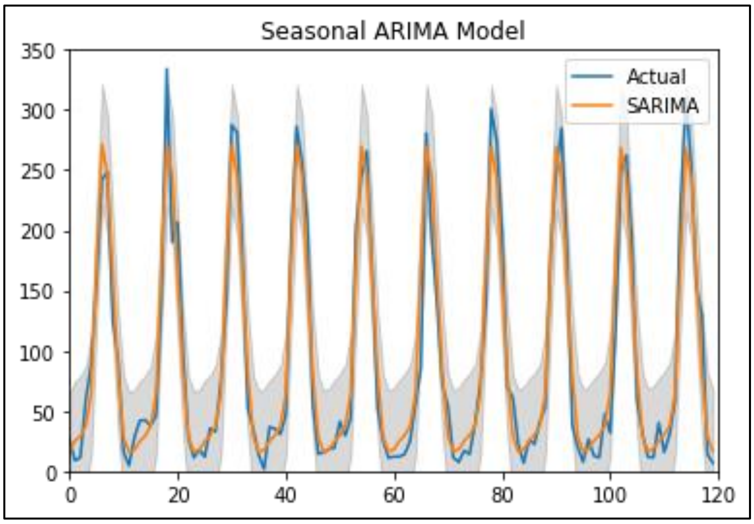

Figure 4.13: SRIMA Classification Results.

Mean absolute error of Seasonal ARIMA model is about 18mm, which is slightly worse than the Naive model.

Recurrent Neural Network (RNN) is widely used for time series prediction. Particularly a variant of RNN called LSTM is vastly popular. Here I have used a simple architecture (with not much tuning) involving 1 LSTM layer and 2 feed forward networks. More details can be found in model summary. Time series generator function is used to set up the training and test data. Look back period is 12 months.

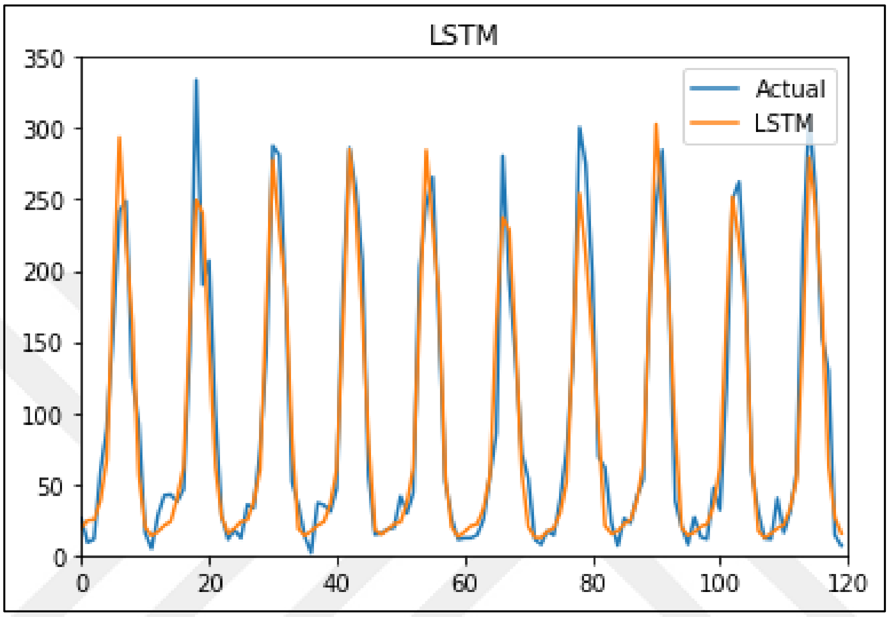

Figure 4.14: LSTM Classification Results.

LSTM also under performs the naive model. Perhaps further tuning of the model could help boost its performance. But I guess the past values of the rainfall is not a good indicator for the future. This is one of the case where a common sense based naive model performs better than the others.

4.5.3 Recognizing the Month by Outgoing Longwave Radiation (OLR)

Outgoing Long-wave Radiation (OLR) is electromagnetic radiation with wavelengths ranging from 3 to 100 m that is emitted from Earth and its atmosphere into space as thermal radiation. It is a measure of the amount of energy radiated into space by the earth's surface, seas, and atmosphere, and it is an important component of the planet's radiation budget.

OLR has a dominant meridional (North - South) annual cycle as it moves from summer to winter and viceversa. OLR is also affected by short term day to day variability as well as long term decadal events. Here I have taken the OLR daily data for 39 years (1979-2017) and have attempted to identify the month to which the data belongs by using a deep neural network.

The OLR data for the 15th of every month is plotted below. From the naked eye its very difficult to differentiate the months. Feel free to change the year and experiment. Download the data and put in the Data folder.

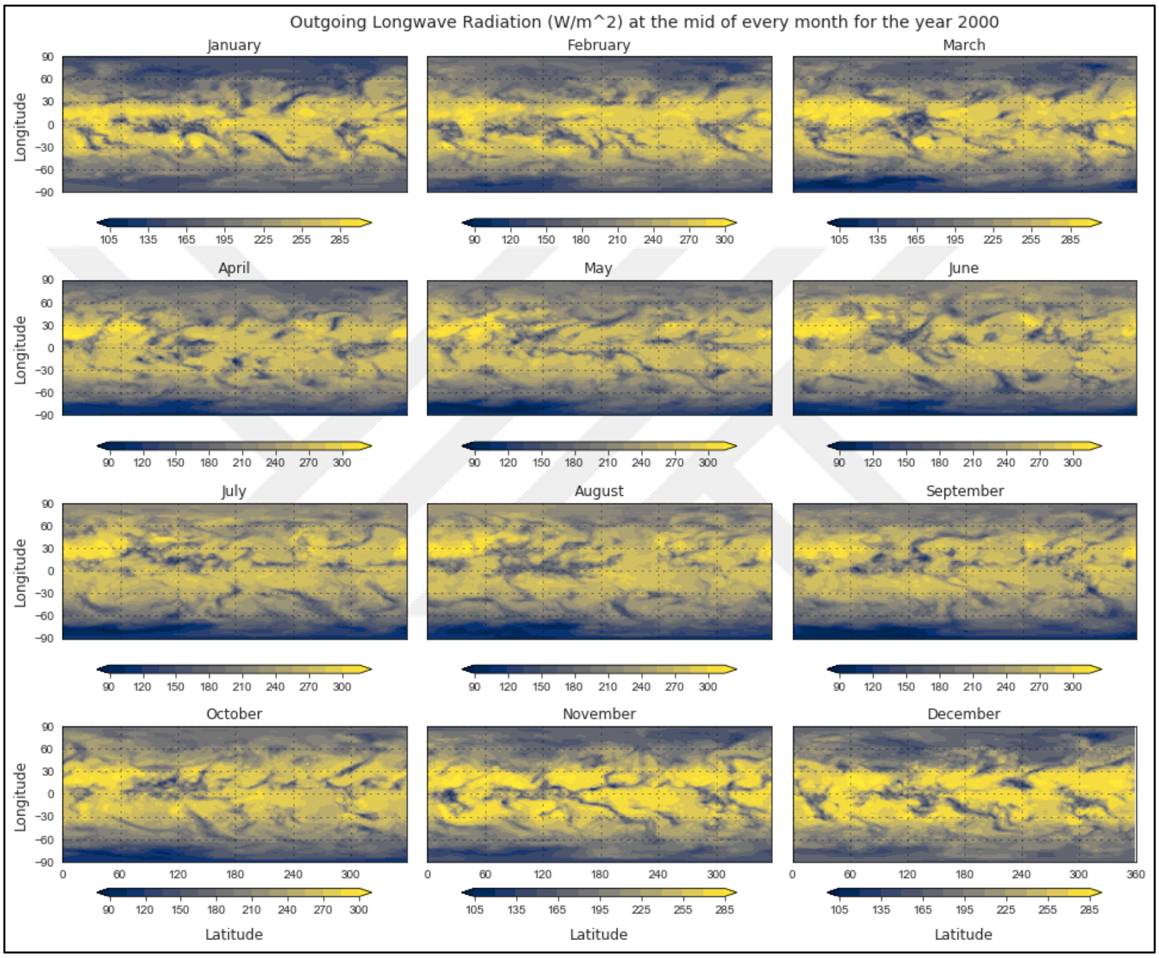

Figure 4.15: OLR Data for The 15th of Every Month.

I have Keras to build a densely connected network. The output (month names) have been transformed by an one hot encoding process. 25% of the data has been randomly selected for testing.

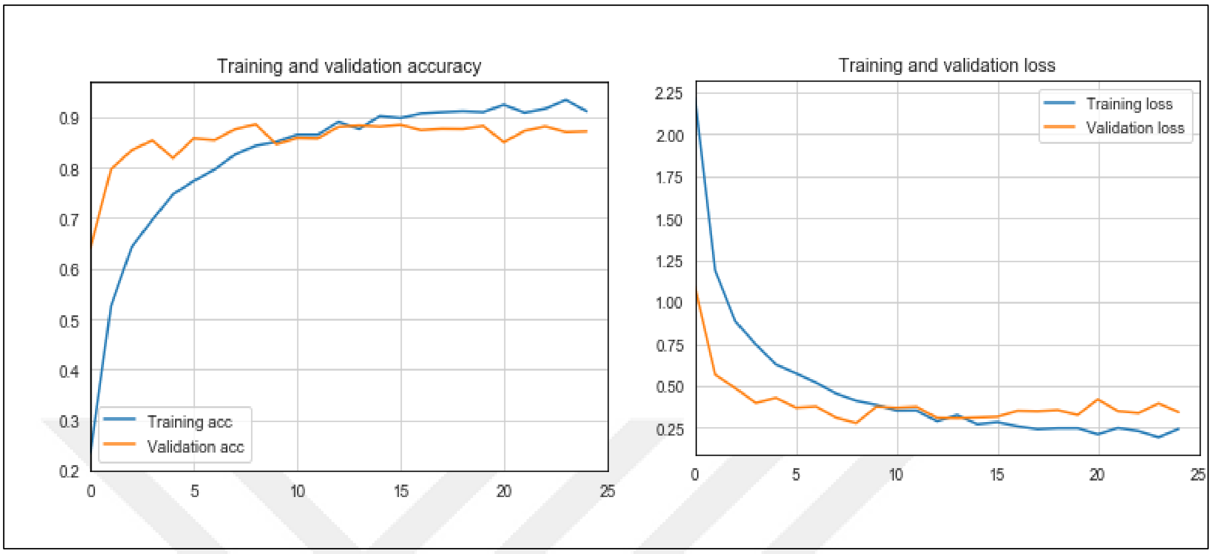

Figure 4.16: Test Data Has an Accuracy.

Test data has an accuracy of about 88%. Now we will have a look at the confusion matrix which gives a better insight of the model performance.

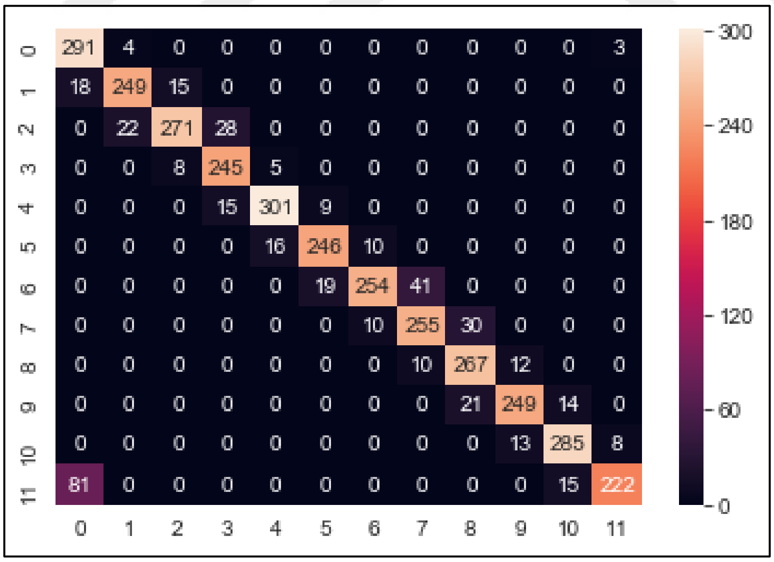

Figure 4.17: Confusion Matrix.

Almost all the misclassification is between the neighboring months. This is expected as for eg. the Jan 30th and Feb 2nd would have similar OLR signature but would be classified as different months. Keeping this in mind we can conclude that the model performs very well in classifying the months.

4.5.4 Retrieving Winds from the Vorticity and Divergence Fields Using ANN

Deep learning is often used to extract features from the data. In geosciences its increasingly used in the parameterization of unresolved sub-grid dynamics. Few examples include the convection parameterization in general circulation model.

In this section we try to retrieve the wind fields from the vorticity and divergence fields. Though its not exactly a parameterization problem but the idea is similar. We use deep learning to learn a function which maps the vorticty and divergence to the wind fields, but without using the spherical transforms. 2.5◦◦ X 2.5◦◦, 6-hourly ERA-Interim data from 1979-2018 (40 years) is used. Vorticity and divergence fields are used as input and zonal and meridional winds are the targets. All the fields are at 500 hpa level.

First 50,000 samples are used for training and the remaining 8440 samples for testing. Sanples are randomly shuffled in each of these sets. Cross-validation is used to evaluate different models. The inputs and outputs are normalized by its training mean and variance, so that they have similar scales. This often helps in speeding up the training.

The model has 2 inputs (vorticity and divergence) and 2 outputs (zonal and meridional wind). Six layer deep, fully connected network is used. First 3 layers are shared between the two outputs. It also has another 3 layers for each of the outputs which is not shared. Each layer has 512 units wuth ReLU activation, except the last one. The last layer has 10512 units (with no activation) which is then reshaped to the output image of size 73 X 144. Adam optimization is used with mean squared loss. No regularization was used as the model was not overfitting. Model summary is printed below along with the block diagram.

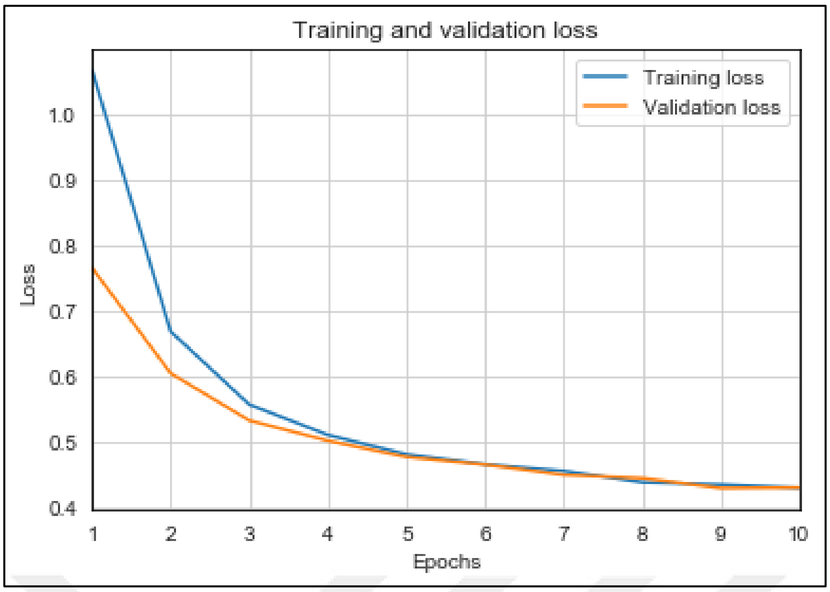

Figure 4.18: Training and Validation Losses.

Training and validation losses are approaching a steady value around epoch 10. Now we will predict the test output.

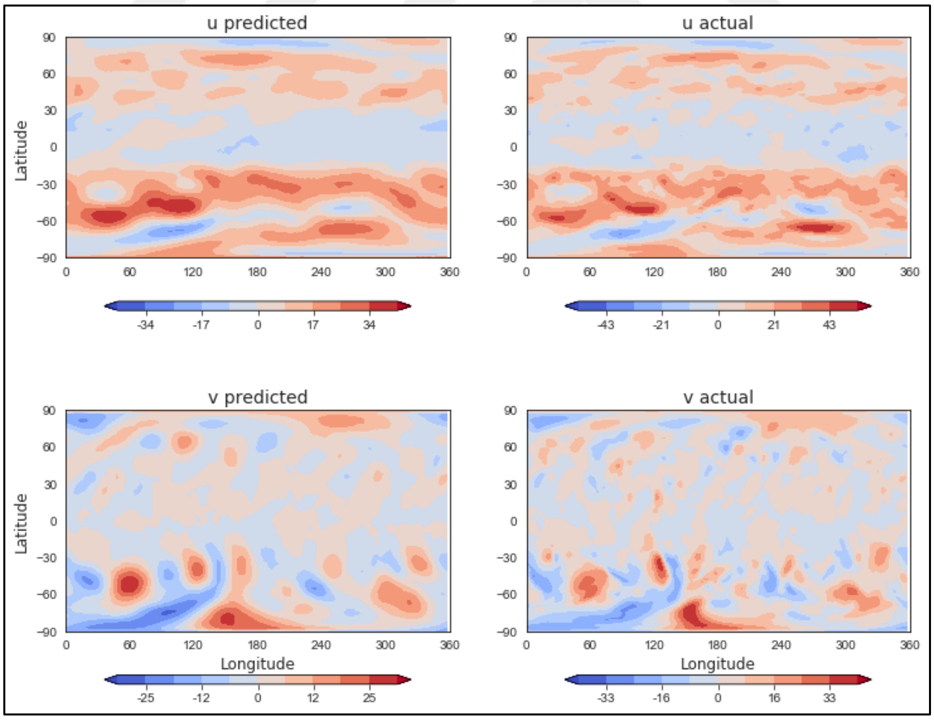

Figure 4.19: Predicted Outputs.

Predicted outputs look more or less similar to the actual ones. Its capturing most of the large scale structures though its more smoothened. The predicted values also underestimate the actual magnitudes. Perhaps a higher capacity network would better capture these subtleties. But overall the performance seems to be satisfactory.

4.6 CONCLUSION

In conclusion, weather prediction is an important task that can be tackled using various machine learning techniques. In this project, we explored the use of Linear Regression, K-Nearest Neighbors, Decision Trees, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks for weather classification and prediction. Linear Regression and K-Nearest Neighbors were used for weather classification based on historical data, while Decision Trees were used for both classification and prediction. CNNs were used to treat weather data as images and predict weather conditions, while LSTM networks were used to predict future weather based on historical data. Each of these techniques has its strengths and limitations, and the choice of which technique to use depends on the specific application and available data. Linear Regression and K-Nearest Neighbors are simple and fast to implement but may not capture complex patterns in the data. Decision Trees are easy to interpret and can handle both categorical and numerical data, but may suffer from overfitting. CNNs are powerful for capturing spatial correlations in weather data, but require significant computational resources and large amounts of data. LSTM networks can capture temporal dependencies in weather data and make accurate predictions, but also require significant computational resources and training data. Overall, the results showed that all of the techniques performed reasonably well for weather classification and prediction, with LSTM networks achieving the highest accuracy in predicting future weather conditions. However, the performance of the models may be impacted by the quality and representativeness of the input data, and further improvements can be made by incorporating additional features or refining the model architectures. In summary, weather prediction is a complex task that can be tackled using a range of machine learning techniques, each with their own strengths and limitations. By carefully selecting and combining these techniques, it is possible to build accurate and reliable weather prediction models that can be used for a variety of applications.

5. CONCLUSION AND FUTURE WORKS

5.1 CONCLUSIONS

In this project, linear regression and deep learning are used to predict the weather forecasting. We divided the whole weather forecasting project into two parts. When compared to the machine learning approach, the deep learning approach produces greater results. Weather forecasts are becoming more accurate and valuable, and their benefits are spreading across the economy. While significant progress has been made in improving weather forecasts, there is still much space for improvement.

For future improvements, following step we thought to took:

a. Replacing model with a latest/different model.
b. Using other robust datasets.
c. Predicting result on more attributes.
d. Training model on higher-end GPU.

Also, while performing weather forecasting, there was a lot of complexities involved. There are a lot of variables/attributes to consider for forecasting weather and if all or most of them are used, then we need a lot of computation power to get weather information. And, Real time weather forecasting is very difficult to forecast correctly.

5.2 FUTURE WORKS

There are two keys areas that comprise promising topics for future inquiry related to the focus of this thesis. Although learning a new RL policy atop an existing policy can combine two exploration methods into one unified policy, achieving the best possible efficiency and performance when training a policy that combines multiple strategies remains an open problem. Future work may benefit from considering the use of curricula in reinforcement learning methods; [72] are able to help train a policy for multi-task purposes. Moreover, the Asynchronous Advantage Actor Critic (A3C) [28] reinforcement learning model can train a multi-task policy similar to ours via parallel training with multiple actors.

REFERENCES

[1] Agilan, V & Nanduri, U. (2016). Modelling nonlinear trend for developing non-stationary rainfall intensity–duration–frequency curve. International Journal of Climatology. 37. 10.1002/joc.4774.

[2] Ali M. F., Asklany S. A., El-wahab M. A., & Hassan M.A. (2019). Data Mining Algorithms for Weather Forecast Phenomena Comparative Study. International Journal of Computer Science and Network Security, 19(9), 76-81.

[3] C. L. P. Chen & C.-Y. Zhang (2014). Data-intensive applications challenges techniques and technologies: A survey on big data. Inf. Sci., vol. 275, pp. 314-347.

[4] Camacho, J., Rodriquez-Gomez, R. A., & Saccenti, E. (2017). Group-wise Principal Component Analysis for Exploratory Data Analysis. Journal of Computational and Graphical Statistics, 26(3), pp. 501-512, doi: 10.1080/10618600.2016.1265527

[5] Cheng, X., Jing, W., Song, X., & Lu, Z. (2019). 5th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2019, Guilin, China, September 20–23, 2019, Proceedings, Part I. Pages 217-218. ISBN: 978-981-15-0117-3

[6] Darji, M. P., Dabhi, V. K., and Prajapati, H. B. (2015). Rainfall forecasting using neural network: A survey, 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, pp. 706-713, doi: 10.1109/ICACEA.2015.7164782.

[7] Fahey, M. (2016, Aug 25). This weather app will give you the most accurate forecasts. CNBC. Retrieved from https://www.cnbc.com/2016/08/25/this-weather-app-will-give-you-themost-accurate-forecasts.html .

[8] Gantz, J.& Reinsel, D. (2012). The digital universe in 2020: Big data bigger digital shadows and biggest growth in the far east. International Data Corporation

[9] Goger, B., Rotach, M., Gohm, A., Stiperski, I. & Fuhrer, O. (2016). Current Challenges Representation and Performance Computing & Simulation (HPCS), Innsbruck, 2016, pp. 890-894, doi: 10.1109/HPCSim.2016.7568428.

[10] Golchha N. (2015). Big data - The information revolution. International Journal of Applied Research. 1, 791-794.

[11] Han, J., Kamber M. & Pei, J. (2000). Data Mining: Concepts and Techniques. Morgan & Kaufmann. ISBN: 9780123814807

[12] Hemalata, P. (2013). Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System. International Journal of Computational Engineering Research 3(3)

[13] International Data Corporation (2019). The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS45213219

[14] Javed, F., Afzal, M. K., Sharif, M., & Kim, B.(2018). Internet of Things (IoT) Operating Systems Support, Networking Technologies, Applications, and Challenges: A Comparative Review. IEEE Communications Surveys & Tutorials, 20(3), pp. 2062-2100, doi: 10.1109/COMST.2018.2817685.

[15] Jolliffe I.T. & Cadima J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374 (2065), doi:10.1098/rsta.2015.0202.

[16] Juneja, A. & Das, N.N. (2019). Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, pp. 559-563. doi: 10.1109/COMITCon.2019.8862267

[17] Kalyankar, M.A. & Alaspurkar, S.J. (2013). Data Mining Technique to Analyze the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), 114-118.

[18] Kumar R. (2013). Decision Tree for the Weather Forecasting. International Journal of Computer Applications 76(2):31-34.

[19] Lin N., Noe D., & He X. (2006). Tree-Based Methods and Their Applications. Springer Handbook of Engineering Statistics. Springer Handbooks. Springer, London. doi:10.1007/978-1- 84628-288-1_30

[20] Mahmood M. R., Patra R. K., Raja R., and Sinha G.R. (2019). A Novel Approach for Weather Prediction Using Forecast Analysis and Data Mining Techniques. Innovations in Electronics and Communication Engineering, 479-489.

[21] Marjani, M. et al. (2017) Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. IEEE Access, Vol. 5, pp. 5247-5261, doi: 10.1109/ACCESS.2017.2689040

[22] Ma, L.-M. Research Progress on China typhoon numerical prediction models and associated major techniques. Prog. Geophys. 2014, 29, 1013–1022.

[23] Tropical Cyclone Forecast Model. Available online: https://en.wikipedia.org/wiki/Tropical_cyclone_ forecast_model (accessed on 1 May 2020).

[24] Moon, I.J.; Ginis, I.; Hara, T.; Thomas, B. A Physics-Based Parameterization of Air-Sea Momentum Flux at High Wind Speeds and Its Impact on Hurricane Intensity Predictions. Mon. Weather Rev. 2007, 135, 2869–2878.

[25] Bender, M.A.; Ginis, I.; Kurihara, Y.J. Numerical simulations of tropical cyclone-ocean interaction with a high-resolution coupled model. J. Geophys. Res. Atmos. 1993, 98, 23245–23263

[26] Lee, C.Y.; Tippett, M.K.; Camargo, S.J.; Sobel, A.H. Probabilistic Multiple Linear Regression Modeling for Tropical Cyclone Intensity. Mon. Wea. Rev. 2015, 143, 933–954.

[27] Wang, Y.; Rao, Y.; Tan, Z.M.; Schönemann, D. A Statistical Analysis of the Effects of Vertical Wind Shear on Tropical Cyclone Intensity Change over the Western North Pacific. Mon. Weather Rev. 2015, 143, 3434–3453.

[28] DeMaria, M.; Kaplan, J. An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and eastern North Pacific basins. Weather Forecast. 1999, 14, 326–337.

[29] DeMaria, M.; Kaplan, J.J.W. A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. Weather Forecast. 1994, 9, 209–220.

[30] Demaria, M.; Mainelli, M.; Shay, L.K.; Knaff, J.A.; Kaplan, J. Further Improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). Weather Forecast. 2005, 20, 531–543

[31] Kim, Y.D.; Choi, S. Nonnegative tucker decomposition. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1–8.

[32] Lipton, Z.C.; Berkowitz, J.; Elkan, C. A critical review of recurrent neural networks for sequence learning. arXiv 2015, arXiv:1506.00019.

[33] Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 2017, 5, 8–36.

[34] Krasnopolsky, V.M. The Application of Neural Networks in the Earth System Sciences. Neural Networks Emulations for Complex Multidimensional Mappings; Springer: Berlin, Germany, 2013

[35] Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the 28th International Conference on Neural Information Processing Systems; MIT Press: Montreal, QC, Canada, 2015; pp. 802–810.

[36] Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Deep learning for precipitation nowcasting: A benchmark and a new model. In Proceedings of the 31st International Conference on Neural Information Processing Systems; Curran Associates Inc.: Long Beach, CA, USA, 2017; pp. 5622–5632.

[37] Zhang, P.; Jia, Y.; Gao, J.; Song, W.; Leung, H.K. Short-term rainfall forecasting using multi-layer perceptron. IEEE Trans. Big Data 2018, 6, 93–106.

[38] Zhu, X.; Li, J.; Zhu, M.; Jiang, Z.; Li, Y. An evaporation duct height prediction method based on deep learning. IEEE Geosci. Remote Sens. Lett. 2018, 15, 1307–1311.

[39] Huang, D.; Du, Y.; He, Q.; Song, W.; Liotta, A. Deepeddy: A simple deep architecture for mesoscale oceanic eddy detection in sar images. In Proceedings of the 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), Calabria, Italy, 16–18 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 673–678.

[40] Lguensat, R.; Sun

[41] M.; Fablet, R.; Tandeo, P.; Mason, E.; Chen, G. EddyNet: A deep neural network for pixel-wise classification of oceanic eddies. In IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Piscataway, NJ, USA, 2018; pp. 1764–1767.

[42] Du, Y.; Song, W.; He, Q.; Huang, D.; Liotta, A.; Su, C. Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection. Inf. Fusion 2019, 49, 89–99.

[43] Wang, Z. What is the Key Feature of Convection Leading up to Tropical Cyclone Formation? J. Atmos. Sci. 2018, 75, 1609–1629.

[44] Halperin, D.J.; Fuelberg, H.E.; Hart, R.E.; Cossuth, J.H.; Sura, P.; Pasch, R.J. An Evaluation of Tropical Cyclone Genesis Forecasts from Global Numerical Models. Weather Forecast. 2013, 28, 1423–1445.

[45] Michael, B. Tropical Cyclone Genesis Forecasting and Pre-Genesis Forecasts Report; National Hurricane Center: Miami, FL, USA, 2017.

[46] Zhang, W.; Fu, B. Discriminating developing versus nondeveloping tropical disturbances in the western North Pacific through decision tree analysis. Weather Forecast. 2015, 30, 446–454.

[47] Park, M.S.; Kim, M.; Lee, M.I.; Im, J.; Park, S. Detection of tropical cyclone genesis via quantitative satellite ocean surface wind pattern and intensity analyses using decision trees. Remote Sens. Environ. 2016, 183, 205–214.

[48] Ahijevych, D.; Pinto, J.O.; Williams, J.K.; Steiner, M. Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. Weather Forecast. 2016, 31, 581–599.

[49] Zhang, T.; Lin, W.; Lin, Y.; Zhang, M.; Yu, H.; Cao, K.; Xue, W. Prediction of Tropical Cyclone Genesis from Mesoscale Convective Systems Using Machine Learning. Weather Forecast. 2019, 34, 1035–1049.

[50] Kim, M.; Park, M.S.; Im, J.; Park, S.; Lee, M.I. Machine Learning Approaches for Detecting Tropical Cyclone Formation Using Satellite Data. Remote Sens. 2019, 11, 1195

[51] Matsuoka, D.; Nakano, M.; Sugiyama, D.; Uchida, S. Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model. Prog. Earth Planet. Sci. 2018, 5, 80.

[52] Richman, M.B.; Leslie, L.M.; Ramsay, H.A.; Klotzbach, P.J. Reducing Tropical Cyclone Prediction Errors Using Machine Learning Approaches. Procedia Comput. Sci. 2017, 114, 314–323.

[53] Alemany, S.; Beltran, J.; Perez, A.; Ganzfried, S. Predicting hurricane trajectories using a recurrent neural network. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence; AAAI Press: Honolulu, HI, USA, 2019; Volume 33, pp. 468–475.

[54] Zhang, Y.; Chandra, R.; Gao, J. Cyclone track prediction with matrix neural networks. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.