National College of Ireland

Predictive Analysis on Drought in North America using Deep Learning

MSc Research Project
MSCDA JAN21-A

Sadhvi Rajkumar Dubey
Student ID: 19199350

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sadhvi Dubey
Student ID:	19199350
Programme:	MSCDA
Year:	Jan 2021
Module:	MSc Research Project
Supervisor:	Dr. Vladimir Milosavljevic
Submission Due Date:	16/12/2021
Project Title:	Predictive Analysis on Drought in North America using Deep Learning
Word Count:	6,058
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sadhvi Dubey
Date:	31/01/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Analysis on Drought in North America using Deep Learning

Sadhvi Rajkumar Dubey
19199350

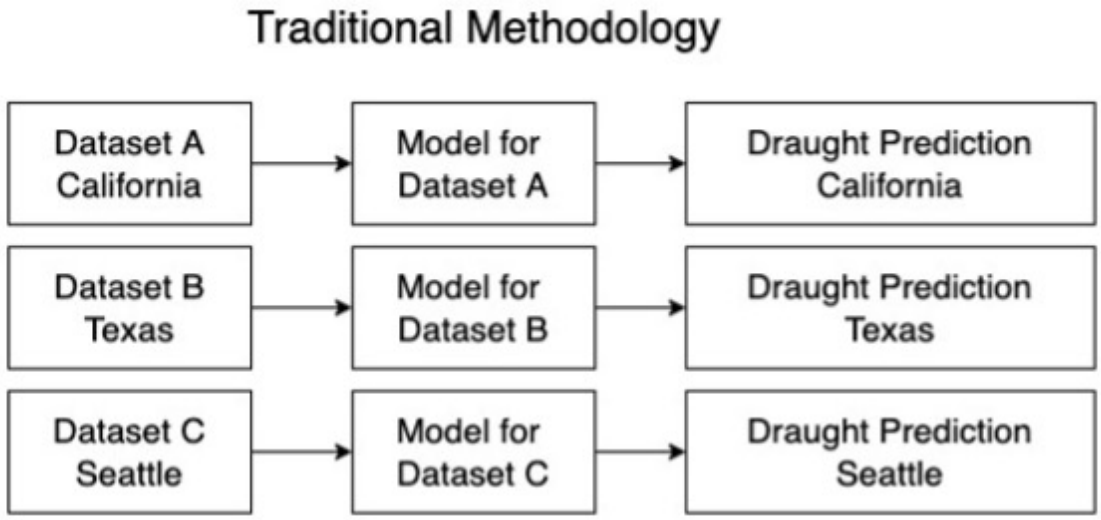
Abstract

The devastating effects of previous drought episodes around the world have prompted considerable drought monitoring and forecast efforts. Various drought information systems with various indicators have already been developed to provide early drought warning. The United States Drought Monitor (USDM), which uses numerous drought categories to classify drought severity and has been used to evaluate and manage by a variety of users such as natural resource managers and authorities, has played a critical part in drought monitoring. The development of drought prediction using USDM drought categories will significantly improve decision-making due to the numerous applications of USDM. This study presented a cross region drought prediction using machine learning model as we can see different climatic condition in different part of United states. To conduct this research we have used USDM dataset which will provide us reliable data from 2000 to present date. The results of USDM drought classification forecasts in the United States show the system's potential, which is projected to contribute to operational early drought warning in the United States. In this paper we put forward a novel cross region drought predicting model in which we will do comparative study using statistical model and machine learning model which will intern provide us most accurate drought prediction. This model can also be used for predicting drought for any country. In comparison to ANN and KNN-based models, LSTM-based models were able to capture the temporal and spatial properties of droughts over the United States better in validation. KNN, which was used for the first time in constructing drought models, performed badly as compared to LSTM and ANN-based drought models.

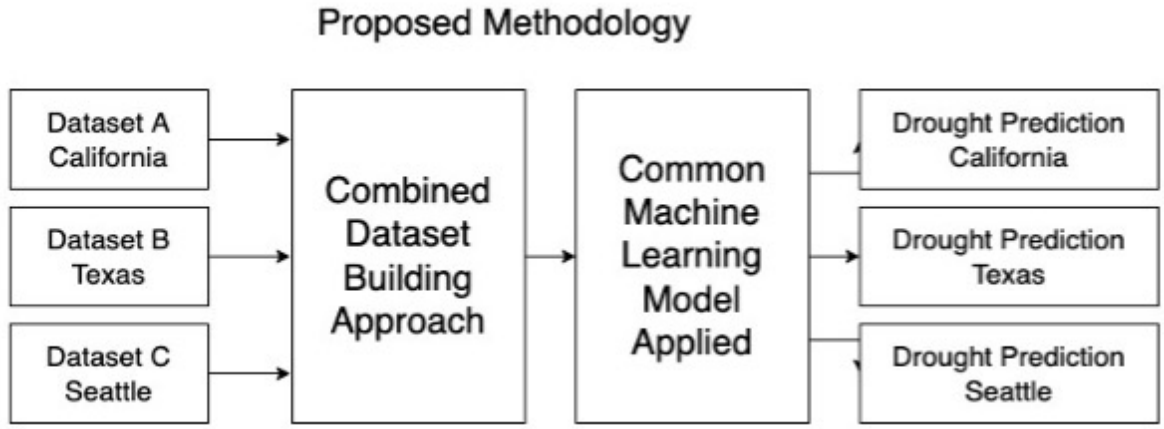
1 Introduction

Since droughts are directly related to the availability of water, their changing characteristics will have profound effects on water stress and food due to climate change. Droughts and their possible occurrence in any geographic area underscore their catastrophic nature among all natural hazards. **Can Machine Learning be applied to predict drought across different regions ?**. The primary novelty of this research lies within the methodology of this research, which is the combination of the dataset building methodology and the models applied to the dataset. Traditional approach towards a drought prediction model involves the usage of a public dataset which is restricted to a particular region in the United States Of America, this technique is beneficial in predicting drought for a

particular region as the model applied to the dataset is specific to that particular dataset.



The model used for one dataset in this case cannot be applied to datasets of the other region as they have a different set of features and value range. Due to this, the model that is trained for one particular region/state cannot be applied to another region/state in The United States Of America.



In this research, the focus is to improve the entire model by introducing a dataset building methodology as well as providing appropriate machine learning models for the dataset. As a primary step, the dataset of several regions is combined based on the publicly available data provided by the government authorities like NASA and NOAA. After the data collected, certain specific regions are selected and combined in order to obtain a more generalised dataset which is applicable throughout The United States. Secondly, the ML models which have showed robust usage in the past have been applied to this

dataset to find out the best model and parameter.This proposed structure can be used to compile data from states of other countries and apply a common model to evaluate the probability of drought. This will help in minimising the data discrepancies found in region specific datasets by laying down a conventional approach towards this problem.

Many droughts have occurred in various parts of the world in the recent past. For example, in East Africa (2010-2011), the drought in Texas (2012), drought in the Central Great Plains of the United States (2012) and drought in California in the USA (2012-2015, the Millennium Drought in Australia (1997-2010)) and the Sahel drought (2012). These major droughts have resulted in immense agricultural losses that have devastated water supplies and crops. A series of severe droughts in US over the past decade, such as the drought in the center of the US in 2012, has made clear the need. Drought early warning to reduce the potential impact. Significant advances have been made in the past decade with various drought forecasting methods from a statistical and dynamic perspective.

Prediction of drought remains a challenge for climatologists and hydrologists due to the complexity of its origin and the size of space-time. Statistical, dynamic, and hybrid models are commonly used to predict drought. In statistical forecasting models, empirical relationships between climate variables and observation-derived drought indicators are used to predict drought. In contrast to statistical models, dynamic models are based on physical interactions between land, sea and atmosphere. These interactions are mathematically represented and resolved in dynamic models to create drought simulations / predictions. Hybrid models, on the other hand, are a combination of statistical and dynamic models. For example, multiple dynamic model forecasts can be combined using a statistical framework that assigns weights to different dynamic model forecasts to derive ensemble forecasts. Due to their simplicity and low computational cost, statistical models are widely used to predict drought.

The main role of Machine learning algorithms in hydrological weather applications is for extreme Natural calamities like Floods and droughts lead to a massive human and economic loss. Machine learning algorithms and models have proven to provide accurate information regarding the time frame and scale of a natural disaster. As the frequency of climate change has increased in the past couple of decades, a more dynamic hydrological weather prediction model is required to give real time and robust information. As the processing powers of CPUs and GPUs have increased by several folds over the last few decades, the real time application of machine learning models is possible. Although after several years of study in this domain, it can be established that there is no one superior method but there are several methods with their separate applications in the field of hydrological weather predictions. As ML models can predict mathematical values by learning from past data, they can be used to predict numerical parameters such as temperature(T), Humidity,Atmospheric pressure and precipitation levels. Using these values, a lot of meteorological predictions can be made to improve response towards dynamic weather patterns.

Machine learning (ML) algorithms are a hard and fast of commands that permit a gadget to routinely research from historic facts and enhance it with out the want for sizable programming. Using distinctive ML algorithms, we expand fashions that may mimic the linear and non-linear interactions among predictors in numerous hydrological

weather applications, consisting of: Precipitation forecasts Precipitation runoff modeling (Yaseen et al., 2015), temperature and warmth wave forecasts (Khan et al., 2019c), drought forecasts. ML algorithms consisting of kNearest Neighbors (KNN), Artificial Neural Network (ANN), Extreme Learning Machine (ELM), Random Forests (RF), Support Vector Machine (SVM), Relevance Vector Machine (RVM), Genetic Programming (GP) Widely used to version complicated interactions among distinctive predictors.

The rationality for selection of the ANN, LSTM and KNN models is because of the draught prediction problem can be classified on a broader scale as a classification problem as it will predict whether the region will face drought conditions or not. The general idea of this study is to identify the most effective model that will suit the dataset in order to set a standard procedure to predict drought using the dataset methodology which is closely based on the Standardised Precipitation and Evapotranspiration Index (SPEI). In the past several studies that have been carried out, the most efficient techniques have turned out to be amongst ANN,LSTM and KNN. Among the three listed techniques, KNN is considered to be one of the oldest ML algorithms and it has been used to predict various statistical and forecasting models. LSTM based models are able to better capture the temporal and spatial characteristics.As this research makes an attempt at establishing a well-structured technique in order to develop a cross-regional drought prediction model, the ideal approach would be to establish baselines using the proven techniques of the past and combining them with the methodology proposed in this project.

2 Related Work

A series of severe droughts in the U.S. in the past decade, such as 2012 central U.S. drought, has highlighted the necessity of early drought warning to reduce the potential impacts. A wide array of regional and global drought information systems has been developed to aid drought early warning(Khan et al. (2020)).

2.1 Drought forecasting system

It can be said that the meteorological drought will be the first step towards a chain reaction.. Although various researchers have advocated to further categorize droughts into different drought types, like(Mishra and Desai (2006)) suggested to include groundwater drought(Beguería et al. (2014)) urged to include environmental droughts and (Li et al. (2019)). This challenge becomes more complicated as it is dependent on who is defining it and what metric is being used for by definition to measure it (**?**). Therefore, drought indicators and indices have been developed, which would provide a general idea about droughts (Sreekesh et al. (2019)). For meteorological droughts, several researchers have defined drought indices based on different variables, and (Yihdego et al. (2019)) provide an extensive list of various drought indices based on drought type. Standard Precipitation Index (SPI) is one of the most used (Hayes et al. (2011)) drought index, one that is arguably the most popular and accepted drought index developed by (McKee et al. (1993)). Compared to other drought indices, it has multiple advantages, like consistent spatial interpretation, less computationally complex, which makes it suitable for prediction and risk analysis (Anshuka et al. (2019)). It is possible to forecast a meteorological drought based on either rainfall or drought indicators (Zhao et al. (2017)). Drought occurrences

are affected by atmospheric circulation patterns or telecommunications (Ganguli and Reddy (2014)). The inclusion of climatic indices to improve forecasting has received conflicting results; in addition, it is well established that future droughts would be affected by climate change. Researchers (Salem et al. (2018))reported that adding climatic indices improves forecasts, whereas (Mariotti et al. (2013)) recorded minimal improvements.

A minimum of 30 years of data is required to understand any drought characteristics. Accurate drought forecasting relies on the length of the available time series, the timescale of the drought index, and the timescale of the model used. This review article explored the capabilities of different drought forecasting models to forecast at different SPI timelines and lead times (forecast range) (Beguería et al. (2014)). Study results showed that one to three months are the best lead time for SPI forecasting. Nonetheless, the determination of SPI timescale is subject to the goals and the dry spell type being examined, alongside the neighborhood highlights, for example, catchment or region type, land-use changes of the review region (Zhao et al. (2017)), with some recommending longer time scales to be utilized and the other proposing the inverse. Their concentrate likewise proposed that SPI at longer time scales are better anticipated as the qualities are smoother and underlines the need to test diverse time scales prior to coordinating it in dry spell the board. Consequently, the current review gauges SPI12 at a lead season of 90 days(Mishra and Desai (2006)).

2.2 Using Machine learning Model drought Forecasting

Because drought is multidimensional, modellers are frequently stumped as to which variables to consider. The incorporation of temporal lag components of climatic variables as predictors in the model was a crucial result for enhancing drought predictions. Several studies have shown that using climatic factors improves the forecasting of drought indices or drought variables. This is especially important for drought studies in Australia, where the relationship between climate drivers and rainfall is the strongest in the world. Because a single measure cannot capture all of the diverse climatic elements, numerous indices are used to evaluate the changes. As a result, numerous indices have been proposed in the literature, and we will use climatic factors throughout this study (pressure indices and SST indices) (Hunt et al. (2018)).

Drought forecasting can be accomplished using statistical, physical, and data-driven methods. The physical-based research entails sophisticated models with multiple variables and a lot of processing power, which is rather difficult. Data-driven models, such as machine learning models, are significantly less intricate than physical-based models and use far less computational resources (Chaudhari et al. (2021)). Although physical-based models have shown equivalent forecasting outcomes, data-driven models are shown to be more effective and can occasionally reach even greater accuracy than physical models. Machine learning (ML) approaches have been used to anticipate droughts and have proven to be particularly successful among data-driven models. Several studies have been undertaken to anticipate drought for various parts of the globe at various lead times using various drought indicators (Ganguli and Reddy (2014))

ML models, on the other hand, subject to the curse of dimensional and over fitting because numerous such factors effect distinct lag times. The deep learning model, on the other hand, is a novel concept that opens up new avenues for boosting predicting capabilities. Several famous academics have said and demonstrated that deep learning would eventually outperform machine learning predicting capabilities and provide a new ap-

proach to regression problems. Over-fitting and dimensional problems can be addressed with LSTMs, where the forget gate determines the amount of data that can flow through. Although LSTM has been employed in a variety of sectors, including banking, meteorological studies, and environmental factors, it is still in its infancy when it comes to forecasting drought indicators and variables. Apart from the desire to anticipate with longer lead times, the interpretation of results in terms of the delays of the variables used to forecast drought is rarely done in drought research. Despite data-driven models' superior performance, evaluating the models and examining the correlations between variables and forecasting predictions, as well as the interrelationships among the variables, is a key difficulty. As a result, the current work aims to create a robust and explainable forecasting model, as well as evaluate the results, in order to better understand how LSTM employs variables to anticipate droughts.

3 Methodology

This section will discuss the techniques used in the implementation of this research, ranging from dataset curating to Data processing. In addition to this, an explanation of the machine learning models applied in the research has been discussed in this section.

3.1 Dataset Curation Methodology:

The dataset in this study is a combination of 180 daily meteorological observations along with the meta-data(to help in plotting) of the locations for a total of 3,108 counties in the United States. This combination of data helps in creating a location-agnostic forecasting model for drought. The input features in this dataset are sourced from mainly the USDM(United States Drought Monitor) Database and also from the NASA climate repository. Observations from previous droughts are also included in the dataset in order to test the models and convey appropriate results.

The USDM data contains drought categories which are in the ranges as shown in the figure below and these categories are renamed according to the correlation in the data provided by NASA. In case of the meteorological data, the model is provided with over 180 observations which leads up to the desired prediction and previous year measurements. The variable is replaced with SPI/SPEI, and for the region which are manually created labels are not available. Additionally, information of season is included in this using sine and cosine. As drought is seasonal, the phenomenon is generalized across season to aid the input.

SPEI Classifications	Categories
≤-2.0	Extremely Dry
−1.99∼-1.5	Severely Dry
−1.49∼-1.0	Moderately Dry
−0.99–0.99	Near Normal
1.0–1.49	Moderately Wet
1.5–1.99	Severely Wet
≥2.0	Extremely Wet

Figure 1: Categories

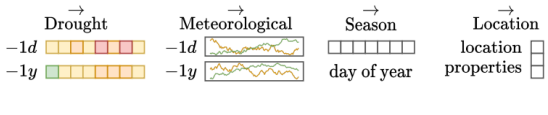


Figure 2: Dataset Curation Methodology.

The dataset encompasses all of the US and each location is indicated with a *location*

indicator,which helps in summarising each location as a combination of parameters from the Harmonized world Soil Database. This dataset is a collection of slope,elevation of each terrain and aspect of the zones.On top of that, it contains the land use of each location.(for example: Canal fed water cultivation or rain fed water cultivation). This location information in terms of soil properties, can enable models to generalize a vast portion of land.

Training data from end portion of are thus helps with another portion of area. The drought vectors in the USDM data is replaced with the SPI/SPEI as well as meteorological vectors to further inculcate temporal data. The location vector in this case is adjusted further by adding binned latitude/ longitude vectors which helps in indicating geographical location of each area. As this dataset is curated from the data available for The United states of America, similar methodology can be applied for another country in the world for which data is available.

3.2 Selection of Climate zones in the Mainland USA:

Mainland USA spans over a coastal boundary of close to 20,000 Km along with vast variety of vegetation and human factors. The figure below illustrates the climatic zones that are present in Mainland US. These zones include Southwest,Gulf Coast/Lower Mississippi Valley/South Atlantic states,Southern Plains/Lower Midwest/Middle East Coast,Northern Great Plains/North-Central/Great Lakes/New England,Pacific Northwest.

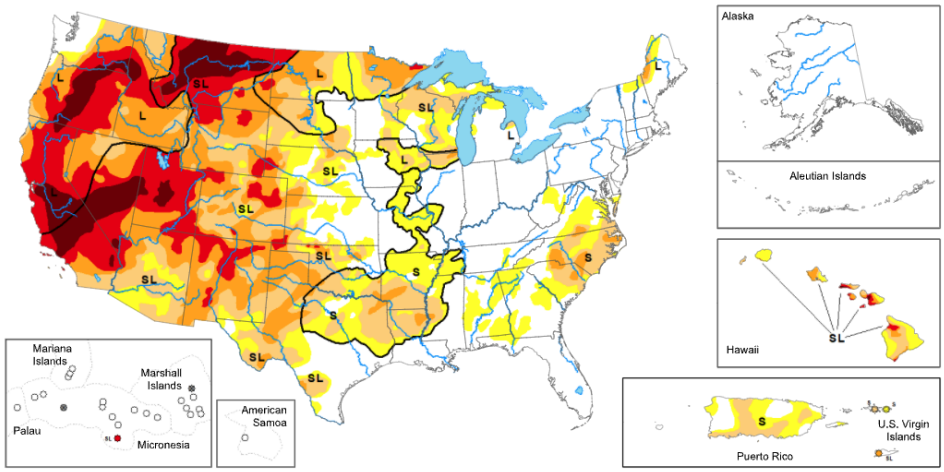


Figure 3: Climatic Zones in the US Mainland

This research does not include the regions of Alaska,Hawaii,Caribbean territories,Pacific territories as the geographical are covered by them is diminished in comparison to Mainland USA. There are 5 climatic zones which are taken into consideration which have a significant impact on the economy and climatic health of the nation.Apart from these 5 climate zones, certain zones are observed to have extreme climatic conditions. One of these zones is the western coast of The United States, which includes the state of California. The state of California is highly prone to droughts throughout the decade which

leads to economic loss and human rehabilitation. The figures below shows the effect of drought in California as categorised by the USDM convention.

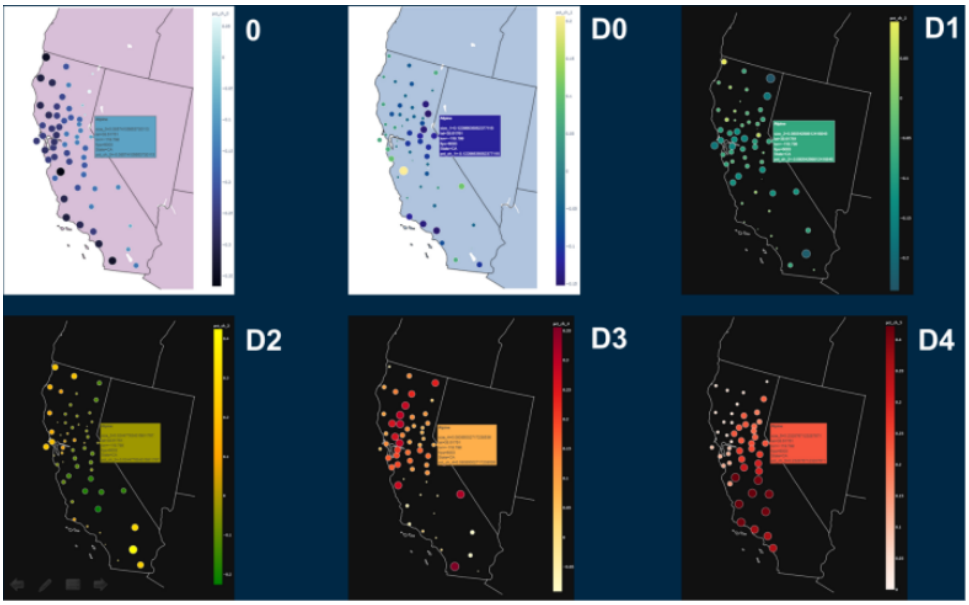


Figure 4: Climatic Zones in the State of California.

It can be observed that California faces severe drought in the all the 5 spectrum of the drought monitor classes. Due to this extreme behaviours of the west coast of The United States of America, it creates an opportunity to use this as a test case in this research.

3.3 Selection of potential predictors:

Potential predictors have been extracted from the probable predictors and correlations between SPEI and PCs which have been derived from the USDM and climate data initially.

Table 1: SPI and USDM category Comparison

Animal	**SPI**	**Description**
D0	-0.5 to -0.7	Abnormally Dry
D1	-0.8 to -1.2	Moderate Drought
D2	-1.3 to -1.5	Severe Drought
D3	-1.6 to -1.9	Extreme Drought
D4	-2.0 or less	Exceptional Drought

The top 1% of the principle components of each probable predictor were identified as potential predictors for drought. These predictors were additional filtered based upon the SPI and USDM category correlation as shown above.

4 Design Specification

This section discusses the design philosophy of the models that are used in the research and how parameters have been tuned in order to achieve best possible results.

4.1 LSTM Model Architecture/Design:

A LSTM model architecture is a special type of Recurrent Neural Network along with a gradient based learning algorithm and it was initially proposed to overcome the error back-flow problems which were faced in traditional RNNs(Sepp and Jürgen, 1997). As shown in the figure below, it is organised as a chain structure and the core of the model lies in the state of the neural networks units which are placed in a series fashion.

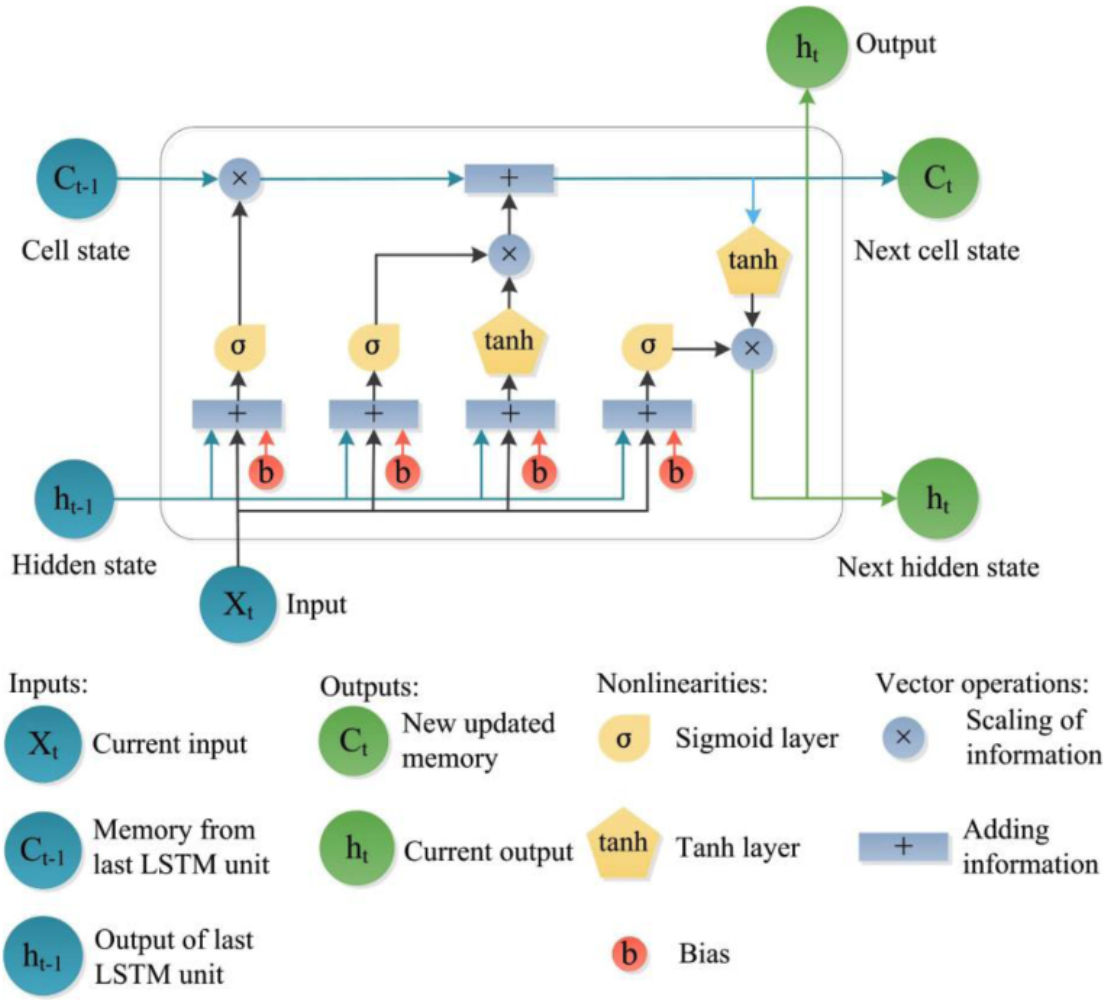


Figure 5: LSTM model architecture

This is transmitted back in a chain structure and flows back into the system. LSTM makes use of three gated controls to operate the cell, these gates control the Cell State(C_t) and the output of the cell (h_t). These gates control the flow of information in the cell and help in avoiding bottlenecks in the models. A gate is similar in structure to a NN layer or a series of matrices that contain different individual weights that carry out do point

multiplication operations.The primary step in this gating mechanism is to make sure the information is forgotten from the sate of the NN unit. Consider the state of the unit to be in C_{t-1}, the previous output h_{t-1} will be read and the new input x_t will be taken into consideration.

$$g_t^f = \sigma(W_f.(h_{t-1}) + g_f)$$

One of the drawbacks of the LSTM model is its inability to process historic data points and to inculcate them wit the current input. Due to this drawback the performance of the LSTM model reduces when the amount of data available is less in number.

Here, σ is the sigmoidal function. In this research, a 5 layer deep neural network has been built in order to increase the yield of the network. The input in this network are the encoded sequences x_1, ... x_n and the Location data is being fed to the FFNN which yields out a 5 stage output which are in correlation to the drought categories defined earlier.

4.2 Transformer Model Architecture/Design:

The architecture of the transformer model is as shown in the figure above. The fundamental of this model is to consume the previously generated output in order to perform operations which help in generating the next symbol.

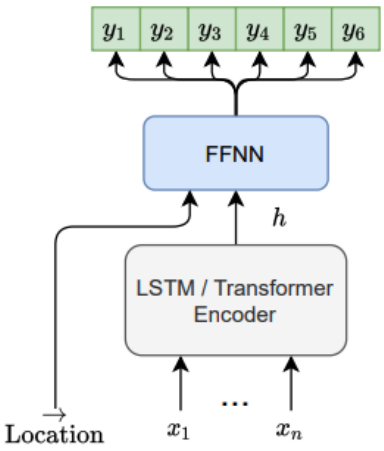


Figure 6: Illustration of passing encoded sequences x_1, ... x_n (combination of meteorological and drought data) along with location vector of FFNN- Feed forward neural network.

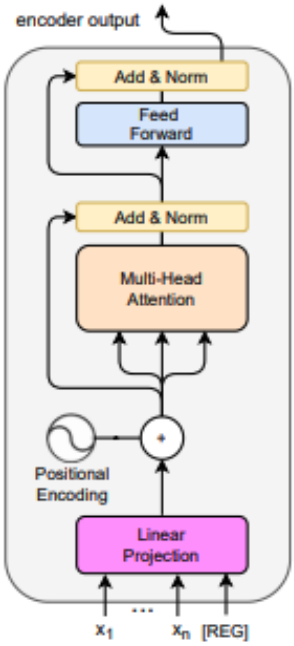


Figure 7: Illustration of the Transformer model along with [REG] token

The transformer model is an improvement on the already existing LSTM model as it adds a multilayered feedback loop in to the system which uses the output of the current step as the input of the next step thereby generating a self propagating loop.In this case, the model is auto-regressive in each step along with a stack of N=5 identical layers. The above diagram represents a self attention mechanism which deferentially weighs the input

data. In the case of this research, this approach is helpful as it makes use of the curated dataset in order to generate an extra set of data points which is used a a test and reference for the current layer and helps with increasing the performance of the overall model. The efficiency of the model is further

5 Implementation

This section gives an overview of the implementation phase of this research in a systematic manner and flow. The first section discusses the Data processing and exploration and the later section discusses the Baselines models applied to assess this data and to generate a comparison.

5.1 Data Exploration and Data Processing

This section explores how data is being explored and processed in this research. In order to establish the usage of the methodology used in this research, it is primarily important to state the relation between the droughts and meteorological data.

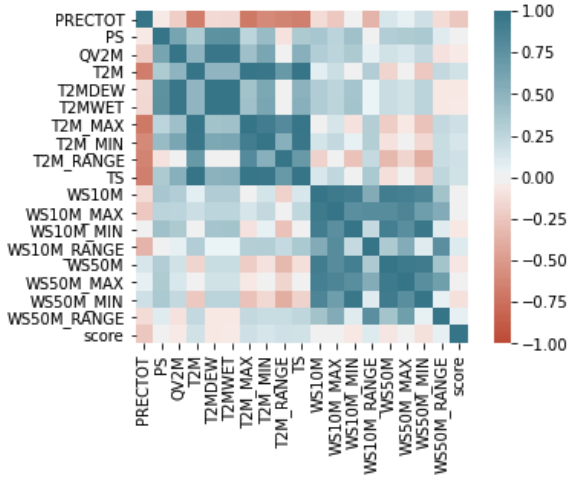


Figure 8: Correlation for Input Features

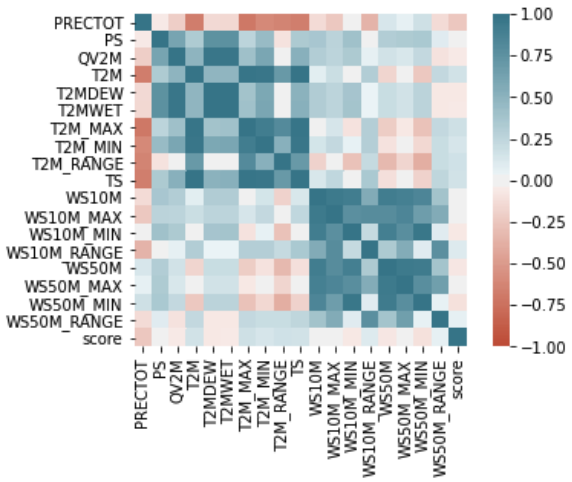


Figure 9: Correlation-Prediction Features

As per this The diagram below shows the correlation between the droughts and meteorological data. On the left the figure shows that the features that are selected are well correlated with the data curated.

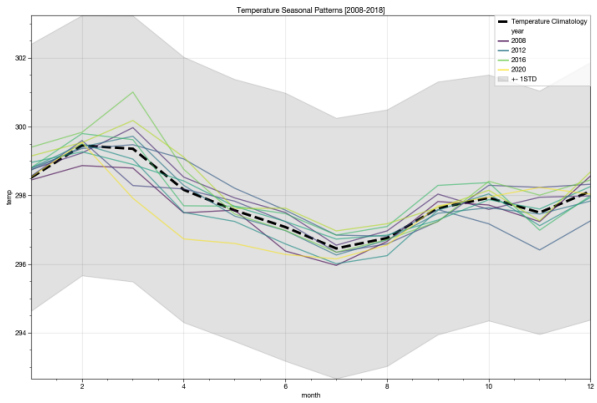


Figure 10: Correlation for Input Features

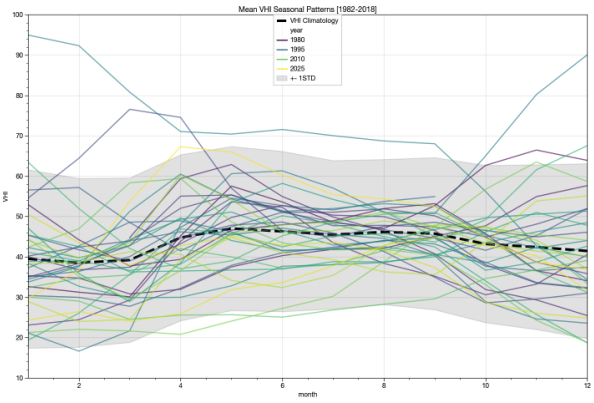


Figure 11: Correlation-Prediction Features

The below diagram shows the availability of data for the entire United states Of America and the Mainland US. This is plotted as per the longitudinal and latitude data available in the dataset.

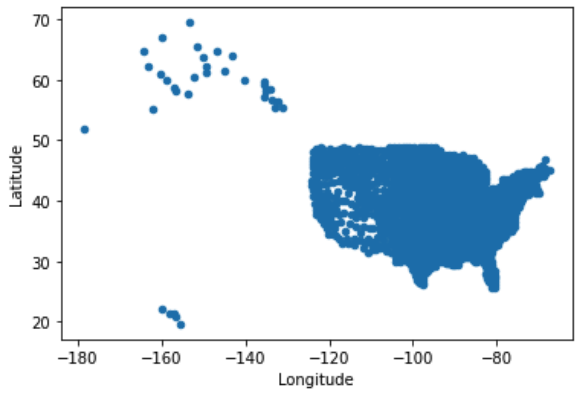


Figure 12: Data point availability in the Entire USA

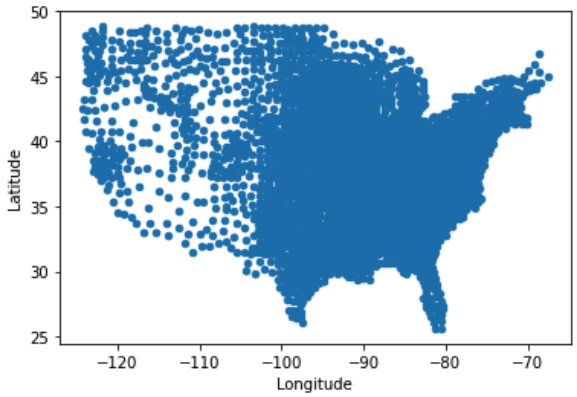


Figure 13: Data point availability in Mainland USA

The figure shows that the density of data points available for the the Central part and Eastern part is more than the data points available for the Western coast of the Mainland.

5.2 Baseline Models(Traditional models currently used for region specific prediction):

The baseline models used for this research are implemented for data with long lead time which is necessary for real world application for generating early warning systems and risk management strategies.The below figure shows the architecture of the baseline model which considers multiple training and testing inputs from the sources and apply a very traditional approach of LSTM and transformer. As a part of this research, this model is applied to the data available for the state of California. This step is performed in order to set baseline and to support the hypothesis of this research.

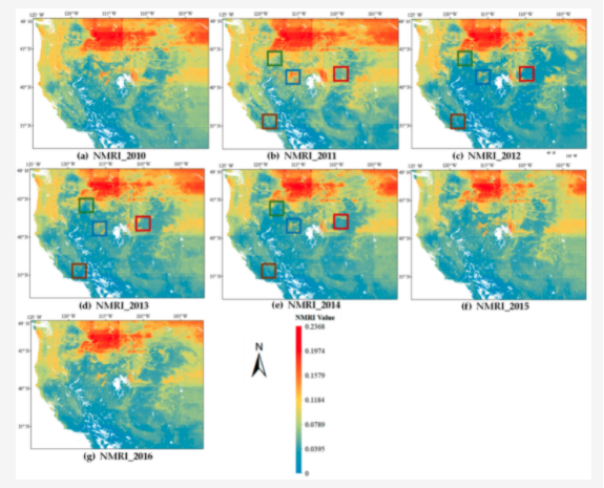


Figure 14: Baseline model to California

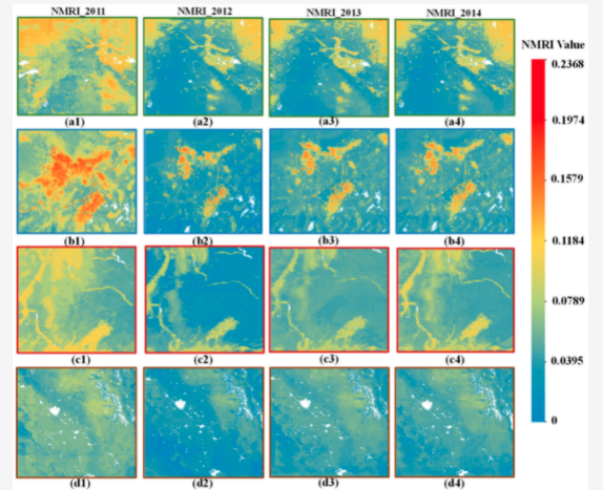


Figure 15: Result Overview

This architecture does not support the dynamicity involved with the dataset of this research as it does not encoder layers in order to deal with the outliers. On the other hand, these types of models work very well for drought prediction of a region specific dataset.As seen from the above results, the drought prediction model applied to the state of California yield good results. In the next step, this model is applied to other states in order to test the cross-region hypothesis proposed in this paper. The image above showcases 4 rows of results out of which three are for the states of Iowa,Montana and Oklahoma. The predicted output for these three states is not as accurate as that of California. The primary reason for this is the region specificity of the model. The model tested here is a region specific model which as been tuned to predict drought in the Californian climatic conditions.

Table 2: Baseline LSTM model results for the State of California

State	Epoch	Results(Average)	
		MAE	F_1
1	California	0.653	62.4
2	California	0.571	68.3
5	California	0.493	70.7
	Iowa	0.095	90.3
All	Montana	0.323	55.8
	Oklahoma	0.181	75.8

The above table shows the performance of the model worsens when it is applied to a different climatic zone. The values of MAE and F1 score in the above table shows that the methodology applied for the model used to predict drought in California does not perform well after being applied to the other states which are in different climatic zones.

6 Evaluation and Results

This section provides an evaluation of the results obtained based on certain factors. In addition to the results, this section also contains graphical representation of the results plotted on the United States Geographical maps. The table shows the Performance for occurrences of a draught conditions.

Table 3: Performance Table for Occurrences

	True not (X)	True label (X)
Predicted label (X)	False Positive (FP)	True Positive (FP)
Predicted not (X)	True Negative (TN)	False Negative (FN)

The output predictions of the models applied in the section below are plotted on the map based on these occurrences which detect the outcome of the model based on the above scale.

6.1 Evaluation Criteria and parameters:

This section discusses the measures used for evaluating the performance of the model. There are two major quantities which are used to evaluate ML models pertaining to the drought prediction domain. The two measures are MAE and F1 score and they are discussed in the sections below:

6.1.1 Mean Absolute Error(MAE):

The below equation represents how the Mean Absolute Error is calculated. Where, $\hat{y}$ is the forecasted value of the drought features and y_i is the original value of the drought features. N is the total number of samples in the dataset.

$$MAE = \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{N}$$

6.1.2 F1 score:

F1 score is measure using which the accuracy of any machine learning model is decided. F1 score is calculated as per the below equations. Mathematically, the F1 score is a harmonic mean of the parameters if the ML model. As per the equation, the parameters here are the precision and recall of the system.

$$F1Score = 2\frac{PrecisionXRecall}{Precision + Recall}$$

Furthermore, this equation can be broken down into the number of False positives,False Negatives, True positives and True negative as shown below:

$$F1Score = \frac{tp}{tp + 0.5(fp + fn)}$$

It be deduced that, precision is the measure of the true positives as compared to the true outputs of the model. Similarly, recall is the measure of sensitivity of the model which is the true positives upon the total of false positives and false negatives in the model.

6.2 Hyperparameter Experiments:

The two models that are applied (I.e. LSTM and transformer model)to the curated dataset is broken down into 4 segments, which are Seasonal Encoding,Location Identifier,Iat/Ion and previous year combination.The below table summarizes the parameters of the LSTM and transformer model used in cross region prediction. Both the models have a batch size of 128 along with a hidden size of 512.

Table 4: Hyperparameter Experiments - Model Setup

Hyperparam	Transformer	LSTM
Number of Layers	4	2
Hidden Size	512	512
Batch Size	128	128
FFNN inner hidden size	4096	N/A
Attention Heads	2	N/A
Initial Projection Size	256	N/A
Dropout Probability	0.1	0.1
Weight Decay	0.01	0.01

Each of the models is then assessed based on the impact of each feature and this is done by retaining the model with other features except the one under observation.

Table 5: LSTM and Transformer model results for Curated Dataset

Models	Week 1		Week 2		Week 3	
	MAE	F_1	MAE	F_1	MAE	F_1
LSTM	0.178	**81.2**	0.233	73.3	0.285	**64.9**
+Seasonal Encoding	0.137	82.2	0.237	71.4	**0.266**	61.9
+Location Indentifier	0.140	**81.7**	0.237	74.3	0.243	**62.9**
+Iat/Ion	0.137	**81.2**	0.237	72.3	0.255	**63.9**
+prev. year combination	0.113	81.2	0.242	73.3	**0.265**	**64.9**
Transformer	**0.158**	68.0	0.215	62.9	0.265	60.1
+Seasonal Encoding	0.178	**88.2**	0.237	71.3	0.295	**69.9**
+Location Indentifier	0.178	79.2	0.237	74.3	0.285	65.9
+Iat/Ion	0.178	**81.2**	0.237	72.3	**0.265**	68.9
+prev. year combination	0.178	81.2	0.237	71.6	0.275	**67.9**

The below table summarizes these models along with their feature isolated results. The impact of each of the individual layer is studied and it is identified that the Location Identifier and Seasonal Encoder have the maximum impact on the models. From future work perspective, the individual layers can be tended to. For this research the dataset is curated in order to remove the effect of bias in the model.

6.3 LSTM and Transformer Model for Mainland US

For the curated dataset, LSTM and Transformer models have been implemented in order to form a generalized approach towards detecting draught in any region for which data is available. The reason for doing this is to create robust models as a part of future work.In case of LSTM, there is no significant movements in all features but there are some noticeable observations and improvements in **Seasonal Information by 2.5%**.Similarly, there is an improvement of **7.6% for Location Information** and an improvement of 9.5% for latitude/longitudinal values. There is an improvement of 16.7% after including meteorological data from the transformer model. After combining the

Table 6: LSTM and Transformer model results for Curated Dataset

Models	Week 1		Week 2		Week 3		Week 4	
	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1
LSTM	0.178	**81.2**	0.237	72.3	**0.265**	**63.9**	0.326	59.6
Transformer	**0.158**	67.0	**0.215**	62.9	0.265	60.1	0.331	56.2
Model Avg Ensemble	0.135	83.6	0.198	**72.7**	0.258	65.1	**0.326**	58.4

The success rate of LSTM has been well documented for predicting drought for a particular region. The map below shows the output of the model plotted to predict drought in the month of June and December. The figure on the left is the predictions the month of June and the figure on the right is for the month of December 2020. As observed that the model has performed well in predicting cross-region zones.

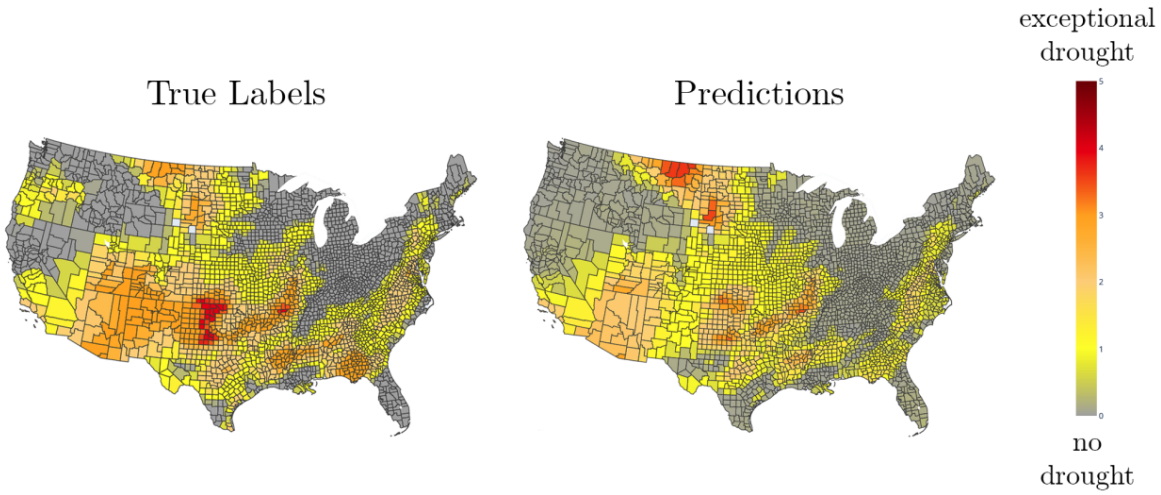


Figure 16: US Mainland - LSTM model prediction

The graphical presentation of the application of Transformer model is plotted below on the map. The predictions observed in the map are far more accurate than the models applied for individual regions.

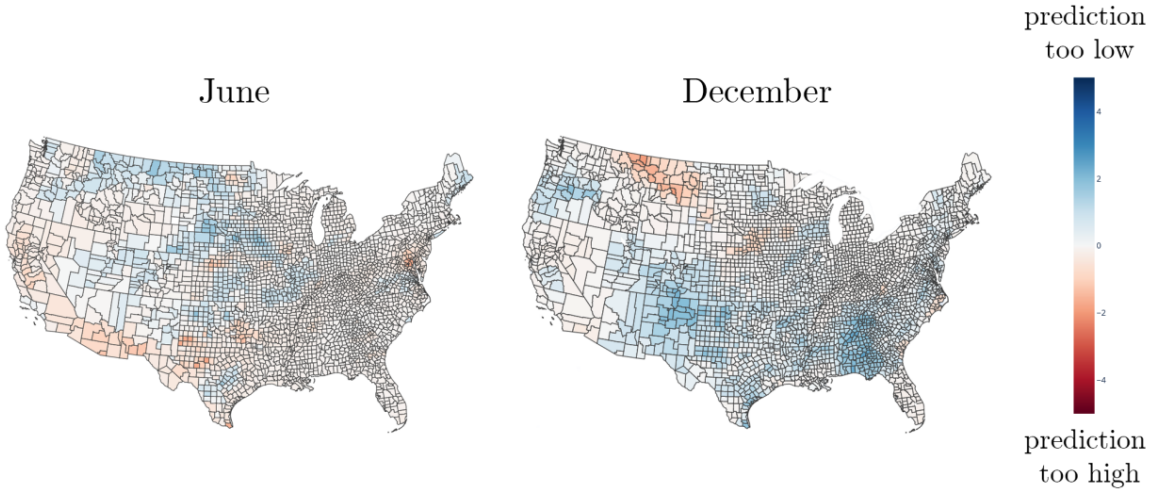


Figure 17: US Mainland - Transformer model prediction

As per the aforementioned research objective, the primary objective is to develop a ML model that is capable enough to predict the probability of drought in a variety of region instead of being a region specific model.This is demonstrated by the two maps as shown in the above figures as the models are capable of handling cross - region data as well as

6.4 Regional Application of the Model:

The below map shows the representation of the selected states on The United States Map for testing the model. The states Iowa,Montana and Oklahoma classify as cross-region states as they span over different climatic and weather zones. The

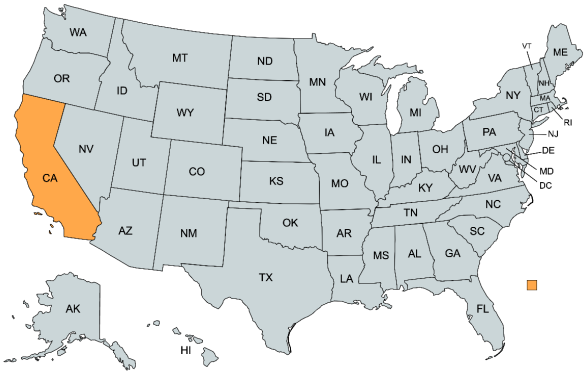


Figure 18: California

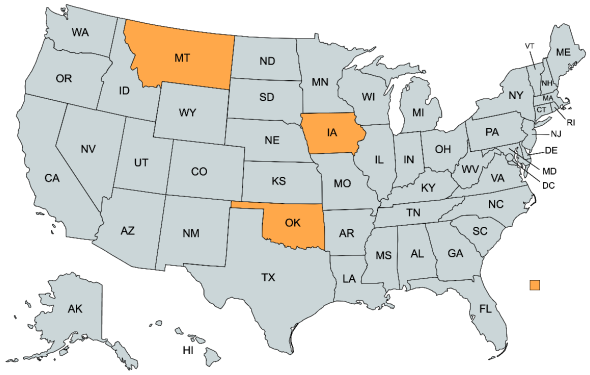


Figure 19: Iowa,Montana and Oklahoma

The table below shows the performance of the applied models on a Local and National level in order to compare the performance of the traditional ML models which are region

specific and the model that is applied in this research which has the aim to be a cross-regional model.

Table 7: LSTM and Transformer model results for Curated Dataset

Training Data	Evaluation Data	Results	
		MAE	F_1
Iowa	Iowa	0.101	88.4
Montana	Montana	0.354	53.3
Oklahoma	Oklahoma	0.213	70.7
All	Iowa	0.095	90.3
	Montana	0.323	55.8
	Oklahoma	0.181	75.8

After training the model on the above mentioned states it is observed that **the model which is trained on data from all the states has an improved performance as compared to the model which his state-specific in nature. The extent of this improvement is as large as 4.6 %**.This is a significant improvement over the state-specific and traditional models.

7 Discussion and Application

In order to demonstrate the significance of this research, an application approach has been discussed in this section. The predictions provided by the model to predict drought can be used in number of ways ranging from Advance warnings to the people of that region and advanced rehabilitation if necessary. One such application is the correlation between draught and forest fires in The United States. The below map shows the correlation between drought and wildfires in the Mainland Unites States.

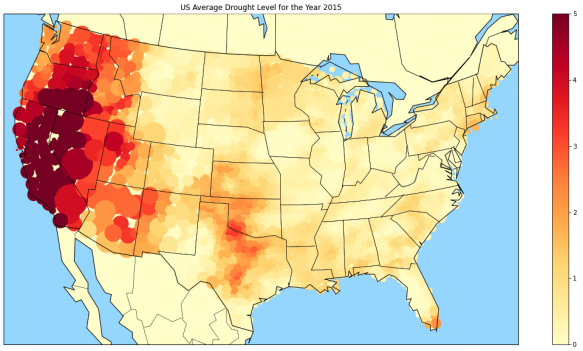


Figure 20: Droughts for Year 2015

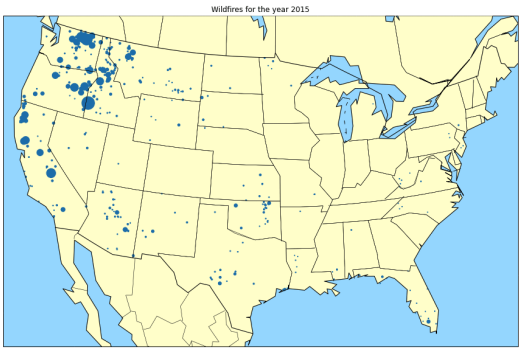


Figure 21: Wildfires for Year 2015

The above figure shows the Drought-Wildfire Plot on the map for the year 2015.The below figure shows the Drought-Wildfire Plot on the map for the year 2016.

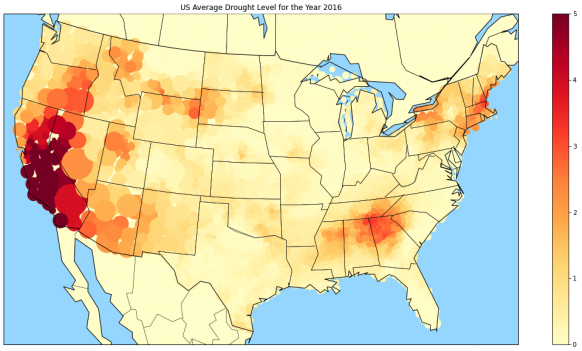


Figure 22: Droughts for Year 2016

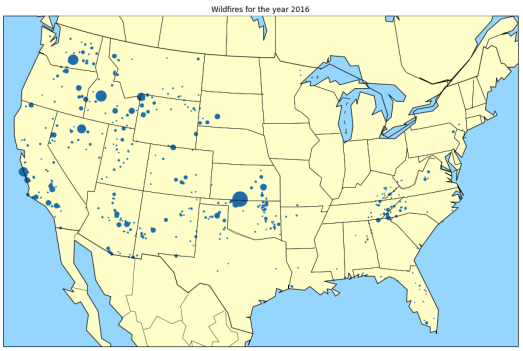


Figure 23: Wildfires for Year 2016

The primary issue faced because of using 180 days of data to model a seasonal effect was the absence of more data points in the training and testing set. This was taken care of by trying several data value pairs and distributing these 180 days of data into two parts ranging from a couple of weeks to a couple of months.As this dataset is a combination of data from multiple sources, the drought season vectors and meteorological data is combined and passed as a sequential inputs to the model. This encoding of the inputs is then used as an input to the single-layer feed forward neural network . This FFNN takes into consideration the location vector and the final output of this network is 6 weeks of drought predictions.By taking this approach, the 180 days of data can be accurately used to predict drought over the period of 6 weeks and testing the model with the help of 18 weeks of data.

By observing the nature of correlation, it can be concluded that there is close relation between the drought prone regions and the regions with wild fire occurrences. Similarly, these predictions can be used on a much more smaller scale which runs the simulation on a county to predict futures wild fires and deploy government resources responsibly.

8 Conclusion and Future Work

The models that are developed in this research are aimed at predicting drought prone regions in order to avoid catastrophic events, human and financial loss. The LSTM and transformer models used in this research efficiently predict the drought prone regions in the Mainland USA as compared to the traditional models implemented in for a geographic and climate specific approach.

Future work can be conducted in the domain of meteorological and location data. The impact of the individual components can be surveyed based upon their weights in the data collection. In addition, a similar methodology can be used in preparing datasets for countries other than The United States Of America. This can be done by collecting NOAA(National Oceanic and Atmospheric Administration) data as well as data available from global drought indicators.

9 Acknowledgement

I am highly grateful to my thesis supervisor Dr. Vladimir Milosavljevic for conducting insightful discussion every week which has immensely helped me in my research project. He has helped me to understand detailed aspect of my project which otherwise I would have lost. I am very thankful for his guidance throughout as it has helped me to complete my project timely.In addition to this, I would also like to thank the National College Of Ireland and other professors who have provided continuous guidance and shared their knowledge which has helped me to grow as an individual.

References

Anshuka, A., van Ogtrop, F. F. and Vervoort, R. W. (2019). Drought forecasting through statistical models using standardised precipitation index: a systematic review and meta-regression analysis, *Natural Hazards* **97**(2): 955–977.

Beguería, S., Vicente-Serrano, S. M., Reig, F. and Latorre, B. (2014). Standardized precipitation evapotranspiration index (spei) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring, *International journal of climatology* **34**(10): 3001–3023.

Chaudhari, S., Sardar, V., Rahul, D., Chandan, M., Shivakale, M. S. and Harini, K. (2021). Performance analysis of cnn, alexnet and vggnet models for drought prediction using satellite images, *2021 Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, pp. 1–6.

Ganguli, P. and Reddy, M. J. (2014). Ensemble prediction of regional droughts using climate inputs and the svm–copula approach, *Hydrological Processes* **28**(19): 4989–5009.

Hayes, M., Svoboda, M., Wall, N. and Widhalm, M. (2011). The lincoln declaration on drought indices: universal meteorological drought index recommended, *Bulletin of the American Meteorological Society* **92**(4): 485–488.

Hunt, K. M., Turner, A. G. and Shaffrey, L. C. (2018). The evolution, seasonality and impacts of western disturbances, *Quarterly Journal of the Royal Meteorological Society* **144**(710): 278–290.

Khan, N., Sachindra, D., Shahid, S., Ahmed, K., Shiru, M. S. and Nawaz, N. (2020). Prediction of droughts over pakistan using machine learning algorithms, *Advances in Water Resources* **139**: 103562.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X. and Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, *Advances in Neural Information Processing Systems* **32**: 5243–5253.

Mariotti, A., Schubert, S., Mo, K., Peters-Lidard, C., Wood, A., Pulwarty, R., Huang, J. and Barrie, D. (2013). Advancing drought understanding, monitoring, and prediction, *Bulletin of the American Meteorological Society* **94**(12): ES186–ES188.

McKee, T. B., Doesken, N. J., Kleist, J. et al. (1993). The relationship of drought frequency and duration to time scales, *Proceedings of the 8th Conference on Applied Climatology*, Vol. 17, Boston, pp. 179–183.

Mishra, A. and Desai, V. (2006). Drought forecasting using feed-forward recursive neural network, *ecological modelling* **198**(1-2): 127–138.

Salem, G. S. A., Kazama, S., Shahid, S. and Dey, N. C. (2018). Groundwater-dependent irrigation costs and benefits for adaptation to global change, *Mitigation and Adaptation Strategies for Global Change* **23**(6): 953–979.

Sreekesh, S., Kaur, N. and Sreerama Naik, S. (2019). Agricultural drought and soil moisture analysis using satellite image based indices., *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* .

Yihdego, Y., Vaheddoost, B. and Al-Weshah, R. A. (2019). Drought indices and indicators revisited, *Arabian Journal of Geosciences* **12**(3): 69.

Zhao, M., Geruo, A., Velicogna, I. and Kimball, J. S. (2017). A global gridded dataset of grace drought severity index for 2002–14: Comparison with pdsi and spei and a case study of the australia millennium drought, *Journal of Hydrometeorology* **18**(8): 2117–2129.