**1.4 Technology and Solution Survey**

Several research papers and journals have investigated drought prediction using multiple datasets and a range of Machine Learning and Deep Learning algorithms. However, due to the absence of a definitive metric for assessing drought severity or percentage, the approaches and models used in these studies have varied. Given the available data resources, we aim to develop a model for forecasting the percentage of drought areas in California using four different Machine Learning and Deep Learning algorithms.

Following are some of the current technologies we have explored for our paper by understanding the reasons behind using those techniques and the limitations of using those models:

**1.4.1 Random Forest Classifier**

Dash et al. (2022) developed a study aimed at predicting drought in various regions of India using the Random Forest Model and Landsat 8 Satellite images. The paper also discusses how water quality will be affected due to drought, with Chlorophyll-a being the primary parameter used to measure this. The study employed a methodology that involved data collection from satellite images, developing a drought prediction model using the extracted data, predicting drought severity, calculating water quality using Sentinel data, and determining the time required for training, along with the model's accuracy. Satellite images were used to create the dataset, with the images being stored as high-resolution .tif files consisting of 11 bands, with each band having a specific .tif file depicting the value of different variables calculated using various formulae, including NDVI, SAVI, MSI, and NIR. Soil Moisture Index (MSI) was used as the output variable to calculate drought severity, with the MSI value ranging from 0 to 1, where 0 represents a severe drought-prone area and 1 represents a non-drought-prone area. A drought model was developed using the Random Forest Model for a total of 1,166,400 images, with the data being split into training (80%) and testing (20%) to determine model accuracy. The model achieved an accuracy of 94% with an error distribution of 5.38% and was trained within 15 minutes. The study also applied the model to predict future drought severity in Sangli town, India, with the SMI value being close to 0.4-0.5, indicating high chances of drought. The paper suggested that water quality is also impacted by the likelihood of drought, which was observed in the Sangli town example.

Masroor et al. (2021) utilized a Random Forest Classifier and Analytical Hierarchy Process (AHP) to assess the impacts of drought conditions on groundwater levels in the Godavari basin. AHP was used to assign weights to the factors and prepare maps for potential groundwater zones. The Standardized Precipitation Index (SPI) was used to measure drought conditions over different periods, and weighted sum overlay analysis was used to identify vulnerable drought zones. The researchers implemented the Random Forest model on this data, considering 80% of the data as training data and 20% of the data as testing data, to assess the impact of drought in those zones. The Random Forest classifier was chosen due to its simplicity and interpretability, as well as its ability to handle missing or outlier data, which is not the case for neural network algorithms. The study found that the Random Forest classifier attained a high accuracy level with an R square value of 0.858. Due to its high accuracy and robustness against overfitting, we are implementing this model as a part of our project.

**1.4.2 Long and Short-Term Memory(LSTM)**

Tian et al. (2021) proposed various models to predict drought severity using feature-based transfer learning and various regression algorithms. In this paper, SPEI is considered as a target feature to predict drought severity and it is considered as four scales such as SPEI-3, SPEI-6, SPEI-9, and SPEI-12. Unlike any other paper, researchers used time series imaging and feature-based technique to extract the features for SPEI value. The main reason to follow this methodology instead of conventional ML or DL-based modeling is there is no perfect metric to find the drought severity and the models become inefficient with changes in the regions and their data. Due to the huge dependency on the study area, researchers have proposed a novel feature extraction based on Time series imaging as their prediction system will be independent of the study area and could be applied anywhere in the world. After feature extraction, this paper used the regression models such as Random Forest, Long and Short-Term Memory(LSTM), Support Vector Regression(SVR), and Wavelet Neural Network(WNN). The main reason to choose Random Forest and Support Vector Regression in this paper is that these models are simplest to use and also consume less time to model the data. LSTM and WNN are neural network models that are used in this paper as these models are good at modeling nonlinear data with a large number of inputs, especially images. When trained with the above models, it is observed that LSTM and WNN models performed extremely well compared to the other two models for long-term periods such as SPEI-9 and SPEI-12. The limitation of this paper is it needed a lot of computational power to work out the paper in real-time as it involves time series imaging technologies and neural networks. Along with this, it is not possible to integrate this model with a conventional dataset approach which will result in the poor performance of LSTM and WNN models.

Dhyani and Pandya (2021) proposed a deep-learning model for drought and agricultural forecasting using Long-Short Term Memory (LSTM) and Time Distributed Convolutional Neural Network (TD-CNN). Satellite images were used to derive values such as NDVI, SPI, and SPEI, and the LSTM model was employed to predict drought occurrence using the SPEI dataset. The study highlighted how different data could impact the performance of each model and recommended predicting SMI and NDVI values using the TD-CNN model since these values are available in the form of images. One of the key advantages of the TD-CNN hybrid model is its ability to recognize sequential relationships between images. Bypassing the extracted features into the LSTM model, which is a specialized Recurrent Neural Networks (RNN) model, the class of the next sequential image can be predicted. After training the hybrid model of LSTM and TD-CNN, the study achieved an accuracy of 96% with a loss of 0.08. The major findings of the study revealed that the hybrid model used SPEI values to build LSTM for drought occurrence prediction, while the TD-CNN model predicted the severity of drought based on NDVI and SMI values. The SMI value was found to be an important metric for determining the type of crop to be planted. One significant limitation of this model is its narrow focus on a limited dataset that is specific to a particular area. As a result, using this model with different sets of data may not yield optimal results.

Mokhtar et al. (2021) conducted a study to develop machine learning and deep learning models to predict drought severity in the Qinghai-Tibet Plateau, which is one of the most sensitive areas to global climatic change. To predict drought severity, the researchers used Extreme Gradient Boost (XGB), Convolutional Neural Network (CNN), Random Forest (RF), and Long-Term Short-Memory (LSTM) algorithms. Since there is no particular attribute to measure drought, the researchers used various scenarios for different variables as datasets. The output variable chosen in this paper is the Standardized Precipitation Evapotranspiration Index (SPEI), which is calculated based on various attributes such as precipitation, temperatures, and relative humidity. The SPEI value is categorized into two sections: the 6-month SPEI value and the 3-month SPEI value. For each category, the climatic data has been divided into 7 cases, with the 7th case consisting of all climatic data including solar radiation, the 6th case consisting of climatic data, and so on. To split the dataset into training and testing, the researchers used PCA and identified five climatic zones, which were used as testing. For the other 25 climatic zones, the aforementioned ML and DL algorithms were performed to predict the SPEI value. The drought severity is measured as follows: SPEI value > 0 is no drought, <-1.65 is very extreme drought, and so on. After analyzing each scenario and category for each model, this paper suggested that the XGB model with scenario 5 is superior in the SPEI-3 category, and the XGB model with scenario 7 is superior in the SPEI-6 category. Among all the models, the XGB model performed the best, producing the highest NSE value of 0.71, followed by the Random Forest model with a value of 0.68. Scenario 7 produced the highest NSE value with the XGB model in the SPEI-3 category. However, in the SPEI-6 category, the RF model produced the highest NSE value of 0.69 in scenario 7 but with an MBE value of -0.08. The major challenge faced by the researchers in predicting drought severity is choosing the correct data, as there is no perfect metric to calculate the drought percentage or severity. This paper demonstrated the usage of various factors in various scenarios to predict drought severity.

Liu et al. (2022) proposed the LSTM model and its variants as algorithms for predicting COVID-19-positive cases in Wuhan and the United States. In this paper, LSTM model prediction is compared with the SEIR model which is an epidemiological model used to predict such contagious diseases and to find the spreading of such diseases. The main reason for choosing LSTM for this paper is that it stores the long-term information in a Cell state such as the previously calculated data and then forgets the information after the cell is updated with other weighted values. In this paper, researchers wanted to use the Bidirectional LSTM model which can store the values in the hidden layer and visible layer which is used for forward and backward propagation respectively. Using this model, researchers were able to achieve an accuracy rate of 94% by training it with the Wuhan dataset and testing it on the dataset from various states in the USA. The Bidirectional LSTM model outperformed the LSTM model and SEIR model in terms of accuracy. The main reason for us to use this model in our project is that it stores long-term information and can handle sequential data to find the patterns and relationships among data which is particularly useful in time series forecasting.

**1.4.3 Artificial Neural Network (ANN)**

Nabipour et al. (2020) presented a model to predict the hydrological drought which affects the water management system using ANN with multiple optimization techniques. This paper analyzed various previous research to understand physical, conceptual, and data-driven models that are used to predict drought percentages. Researchers believed that physical and conceptual-based models are data intensive and could be performed well when there is all the data available to predict. Since there is no concrete data to predict drought, it is better to develop a data-driven model which can perform well with limited data available. The researchers opted to use Artificial Neural Network (ANN) as a model due to its capabilities in parallel processing, handling limited and noisy data, and pattern recognition. Despite its advantages, using ANN has certain disadvantages, as the architecture of ANN is considered based on a trial and error method and it is a black-box model that does not consider the input and output intermediate processes. To address these issues, four optimization techniques are hybridized with ANN which are Grasshopper Optimization Algorithm (GOA), Salp Swarm algorithm (SSA), Biogeography-based optimization (BBO), and Particle Swarm Optimization (PSO). The data is taken from the Dez Dam basin which is used to train the data using the conventional ANN model and hybridized models. After training the data, research showed that the hybridized models performed extremely well compared to the conventional model especially the ANN-PSO model showed great results with an accuracy rate of 84%. The main disadvantage of this paper is that the model takes a lot of time to train the data and huge computational power is required. Along with that, it needs lots of data to get better results.

Agana and Homaifar (2017) proposed a model to predict the long-term approach of drought severity using deep learning techniques such as ANN and Deep Belief Network (DBN). This paper discussed the multiple previous papers on machine learning algorithms that have been used to predict drought severity and came up with a deep learning algorithm of ANN for their model. The main reason to choose this algorithm is that the effects of climate change and factors involved in drought need to be addressed with a much more complicated algorithm ANN, which is suitable for complex time series forecasting and approximating any complex function. Despite having these many layers and complexity to implementing the model, when there are more than two hidden layers needed for highly complicated systems, the non-convex optimization problem arises. To overcome this issue, the Deep Belief Networks (DBN) algorithm has been used to implement the model. The problem that arises with ANN is that the initial values of the weights are randomly initialized, and this may lead to slower convergence or the training model may get stuck at local optima in hidden layers. Hence, the DBN performs greedy unsupervised learning to generate initial values under the dataset used and then applies supervised methods to fine-tune the existing network. This DBN is structured using Restricted Boltzmann Machines (RBM), which are generative energy-based models that are used to configure the hidden and visible layers. By using the feature activations from one layer as the training data for the following layer, the multilayer neural network can be effectively trained. Layer-wise unsupervised training can produce better initial values of weights for all layers than random initialization. The DBN was used to start setting the parameters, then unsupervised learning was used to train the DBN, and the backpropagation algorithm was used to fine-tune the entire network. This paper also used MLP and SVR models to compare the accuracy of the DBN model, which outperformed the other models with 76% of accuracy. The major drawback of the paper is that the data is minimal, and for models like DBN, vast amounts of data are required to improve the performance. The only use of a standardized data flow index could be another drawback. The main reason for us to choose ANN is because of its parallel processing capability and suitability for time-complex problems because of its multilayer functionality which we think will give high performance when trained with our data

**1.4.4 Decision Tree Classifier**

Prasad et al. (2013) proposed a methodology to forecast rainfall by leveraging data mining techniques and a decision tree classifier. The authors selected this algorithm due to its ease of implementation, effectiveness, and compatibility with data mining techniques. The researchers utilized the Supervised Learning in Quest (SLIQ) approach to training the model, which eliminates the need for sorting data by storing it in a separate memory location for each attribute in the dataset. The study yielded an accuracy rate of 72%, demonstrating the model's effectiveness in predicting precipitation, which is a key attribute in rainfall forecasting. However, the authors acknowledge that this approach may perform better with less complex data due to its simplicity and ease of implementation.

Chopda et al. (2018) proposed a technique to identify cotton crop disease by employing a decision tree classifier that considers various parameters such as temperature, soil moisture, precipitation, and more. The authors chose to use the decision tree classification algorithm for its ease of use, as it can be visualized and interpreted to make decisions, which is a primary objective of the study. The paper demonstrates how a decision tree classifier can provide quick and easy predictions of cotton crop disease using basic parameters. We selected the Decision Tree classifier for our study primarily due to its simplicity in interpretation, faster training time, and lower risk of overfitting compared to other neural network algorithms used in this research.

**1.4.5 Comparison of Existing Technology**

According to Dash et al. (2022) and Masroor et al. (2021), both authors proposed research papers to assess drought severity using a Random Forest classifier and satellite images, but their approaches are entirely different. In Dash et al. (2022), sequential images were used as input data for the model, achieving a high accuracy rate of 94%. On the other hand, Masroor et al. (2021) calculated the SPI value from the images and trained the extracted data with the model, achieving an R square value of 0.858. When comparing both papers, Dash et al. (2022) achieved higher accuracy by training the images based on the bands than by extracting SPI values.

The above two papers by Nabipour et al. (2020) and Agana and Homaifar (2017) used ANN as their model in which one of the papers used optimization techniques and the other paper used the DBN algorithm to predict drought. However, each paper used hybrid ANN for modeling and for training the data but the approach followed by Nabipour et al. (2020) is the conventional way of taking the training data from one basin and training the data to calculate the accuracy of the model but the approach followed by Agana and Homaifar (2017) is that they used backpropagation algorithm to fine-tune and train the data. The reason behind choosing hybrid models for each paper is to handle the architecture of the ANN model which is a major challenge in both cases. Furthermore, the accuracy rates of both papers were impacted by the inadequacy and restricted nature of the data, since ANN necessitates an enormous amount of data to train the model and produce better outcomes. It would be inappropriate to evaluate which approach is superior since each approach has its own set of advantages and disadvantages.

**1.5 Literature Survey of Existing Research**

Numerous papers, academic journals and research have been published and disseminated regarding the root causes and consequential impacts of drought, along with the utilization of various machine learning algorithms for accurately predicting such circumstances. Although interpolation techniques based on multiple datasets could potentially yield insight into drought percentages, the utilization of machine learning and deep learning models has significantly improved precision levels. There are many Machine Learning models such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree Classifier, boosted trees, bagging, Matern 5/2 Gaussian process regression (GPR), and M5P Regression. However, obtaining extensive historical data for the particular regions under study is a considerable difficulty for researchers. Based on the available data, many factors were taken into consideration in the papers and research. For instance, many research papers have considered the Standardized Precipitation Index (SPI), Normalized Difference Vegetation Index (NDVI), Atmospherically Resistant Vegetation Index (ARVI), Soil Adjusted Vegetation Index (SAVI), Standardized Precipitation Evapotranspiration Index (SPEI), and other relevant variables when predicting drought percentage. This section discusses the existing research papers used and their results.

Swain et al. (2018) aim to investigate the future trends in precipitation in California using climate model projections. The authors look at how precipitation events have changed in California over the past twenty years in terms of their frequency, intensity, and seasonality. According to the study, while the amount of precipitation in California is expected to rise on average, it will do so with a different distribution, with longer dry spells and more frequent and intense precipitation events. The paper finds that California's precipitation will become more volatile, increasing the risk of droughts and floods and making rainfall patterns less predictable. According to the authors, California's current infrastructure for managing water resources may not be sufficient to handle the rising variability of precipitation, requiring the development of new strategies for water allocation and storage. The study also identifies anthropogenic climate change as the primary cause of the rise in precipitation volatility, which has already increased the likelihood of droughts in California. The paper argues that the government needs to consider the increasing volatility of California's precipitation in its long-term planning for water resources and infrastructure. In conclusion, the paper emphasizes the necessity of proactive planning and adaptation to lessen the effects of future changes in precipitation on California's water resources and economy using any available statistics and technologies.

Wang et al. (2019b) proposed an analysis of drought severity using temporal and spatial variation techniques, using the Standard Precipitation Evapotranspiration Index (SPEI) value as a target metric. According to the study, three kinds of metrics could be used to measure drought severity. They are Standard Precipitation Index (SPI), Palmer Drought Severity Index (PDSI), and Standard Precipitation Evapotranspiration Index (SPEI). The main reason to choose SPEI as the target metric is that the SPI value can only be calculated at different time intervals and will not consider any other variables that could be the factors of drought. PDSI is based on the soil water balance equation, which does not indicate the multi-scalar character. Since SPEI overcomes all these issues, this metric is used as an indicator for drought analysis based on spatial and temporal techniques. This is the reason why we chose the SPEI value as one of the metrics for our project, as it will cover all the factors involved in drought prediction.

Sardar et al. (2022) used a Convolutional Neural Network (CNN) model in combination with the Barnacles Mating Optimizer (BMO) algorithm for predicting drought based on satellite images in Karnataka, India. In this paper, an ensemble hybrid algorithm was used to attain high accuracy for the prediction. The study involved a methodology comprising data collection, computation of drought indices, dataset preparation, ensemble model creation, and performance analysis. Satellite images were sourced from the Indian government's official website and processed into a database object using Rasterio's function from the python library, with multiple values such as NDVI, SAVI, EVI, and ARVI being derived based on the green vegetation color, red color, and blue color. The extracted data is stored in the database once the required values are preprocessed and calculated. After storing in the database, entire data were loaded and divided into training and testing datasets with the proportion of 70 and 30 percent each and then trained using the CNN model. The resulting output was categorized as severe drought, moderate drought, or low drought based on the NDVI value. The accuracy and time taken by the model were used to analyze the model's performance, with an accuracy of 91% achieved using the CNN model alone, but increasing to 94% when combined with the BMO algorithm. The study also suggested that implementing a hybrid model based on different bio-inspired algorithms could further enhance prediction accuracy.

**References**

V. Sardar, S. Chaudhari, A. Anchalia, A. Kakati, A. Paudel and B. B N, "Ensemble Learning with CNN and BMO for Drought Prediction," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/GCAT55367.2022.9971991.

https://ieeexplore.ieee.org/document/9971991

Swain, D.L., Langenbrunner, B., Neelin, J.D. et al. Increasing precipitation volatility in twenty-first-century California. Nature Clim Change 8, 427–433 (2018). https://doi.org/10.1038/s41558-018-0140-y

https://www.nature.com/articles/s41558-018-0140-y

W. Wang, L. Sun, Z. Pei, Y. Chen and X. Zhang, "Analysis of Temporal and Spatial Variation of Growing Season Drought in Jiling Province Based on Standardized Precipitation Evapotranspiration Index," 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Istanbul, Turkey, 2019, pp. 1-5, doi: 10.1109/Agro-Geoinformatics.2019.8820436.

https://ieeexplore.ieee.org/document/8820436

Abbasi, A., Khalili, K., Behmanesh, J. et al. Drought monitoring and prediction using SPEI index and gene expression programming model in the west of Urmia Lake. Theor Appl Climatol 138, 553–567 (2019). https://doi.org/10.1007/s00704-019-02825-9

https://link.springer.com/article/10.1007/s00704-019-02825-9#Sec9

W. Tian, J. Wu, H. Cui and T. Hu, "Drought Prediction Based on Feature-Based Transfer Learning and Time Series Imaging," in IEEE Access, vol. 9, pp. 101454-101468, 2021, doi: 10.1109/ACCESS.2021.3097353.

https://ieeexplore.ieee.org/document/9486838/

N. Nabipour, M. Dehghani, A. Mosavi and S. Shamshirband, "Short-Term Hydrological Drought Forecasting Based on Different Nature-Inspired Optimization Algorithms Hybridized With Artificial Neural Networks," in IEEE Access, vol. 8, pp. 15210-15222, 2020, doi: 10.1109/ACCESS.2020.2964584.

https://ieeexplore.ieee.org/document/8951168

N. A. Agana and A. Homaifar, "A deep learning based approach for long-term drought prediction," SoutheastCon 2017, Concord, NC, USA, 2017, pp. 1-8, doi: 10.1109/SECON.2017.7925314.

https://ieeexplore.ieee.org/document/7925314

Y. Dhyani and R. J. Pandya, "Deep Learning Oriented Satellite Remote Sensing for Drought and Prediction in Agriculture," 2021 IEEE 18th India Council International Conference (INDICON), Guwahati, India, 2021, pp. 1-5, doi: 10.1109/INDICON52576.2021.9691608.

https://ieeexplore.ieee.org/document/9691608

A. Dash, S. Jetley, A. Rege, S. Chopra and R. Sawant, "Drought Prediction and Water Quality Estimation using Satellite Images and Machine Learning," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1110-1116, doi: 10.1109/ICCES54183.2022.9835727.

https://ieeexplore.ieee.org/document/9835727

A. Mokhtar et al., "Estimation of SPEI Meteorological Drought Using Machine Learning Algorithms," in IEEE Access, vol. 9, pp. 65503-65523, 2021, doi: 10.1109/ACCESS.2021.3074305.

https://ieeexplore.ieee.org/document/9408611

S. Liu, Y. Zhou, X. Yang and J. Yin, "COVID-19 positive cases prediction based on LSTM algorithm and its variants," 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), Hangzhou, China, 2022, pp. 268-271, doi: 10.1109/CACML55074.2022.00052.

https://ieeexplore.ieee.org/document/9852599

Masroor, M., Rehman, S., Sajjad, H., Rahaman, M. H., Sahana, M., Ahmed, R., & Singh, R. (2021). Assessing the impact of drought conditions on groundwater potential in Godavari Middle Sub-Basin, India using analytical hierarchy process and random forest machine learning algorithm. Groundwater for Sustainable Development, 13, 100554. https://doi.org/10.1016/j.gsd.2021.100554

https://www.sciencedirect.com/science/article/pii/S2352801X21000102

N. Prasad, P. Kumar and N. Mm, "An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree," 2013 4th International Conference on Intelligent Systems, Modelling and Simulation, Bangkok, Thailand, 2013, pp. 56-60, doi: 10.1109/ISMS.2013.27.

https://ieeexplore.ieee.org/document/6498236

J. Chopda, H. Raveshiya, S. Nakum and V. Nakrani, "Cotton Crop Disease Detection using Decision Tree Classifier," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, India, 2018, pp. 1-5, doi: 10.1109/ICSCET.2018.8537336.

https://ieeexplore.ieee.org/document/8537336