ADDIS ABABA UNIVERSITY
COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

PREDICTING WHEAT YIELD FROM TIME SERIES NDVI VALUES USING A HYBRID ARIMA-BILSTM MODEL WITH ATTENTION MECHANISM

BY
HIWOT TESHOME

ADVISOR
Dr. FITSUM ASSAMNEW

A thesis submitted to the School of Electrical and Computer Engineering in partial fulfillment of the requirements for the Degree of Master of Science in Computer Engineering

June , 2025
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

The undersigned have examined the thesis titled:

PREDICTING WHEAT YIELD FROM TIME SERIES NDVI VALUES USING A HYBRID ARIMA-BILSTM MODEL WITH ATTENTION MECHANISM

BY

HIWOT TESHOME

Approval by Boards of Examiners

Dr. Sosina Mengistu	_____	_____
Dean, SECE, AAiT	Date	Signature
Dr. Fitsum Assamnew	_____	_____
Advisor	Date	Signature
_	_____	_____
Internal Examiner	Date	Signature
_	_____	_____
External Examiner	Date	Signature

Declaration

I, Hiwot Teshome Demissie, declare that this thesis is my original work and that all sources of information in this study have been adequately acknowledged. In addition, I confirm that this thesis has not been submitted in part or in full for any other requirements to any other learning institution.

Student Name: Hiwot Teshome Demissie

Signature: _____

Date: _____

June, 2025

Acknowledgments

I would like to express my deepest gratitude to God for providing me with guidance, strength, and wisdom throughout my life, especially during the course of this research. The divine presence has been a constant source of inspiration, and I am truly thankful for the blessings that have illuminated my journey.

I am also profoundly indebted to my family for their boundless love, unwavering support, and enduring patience. Their encouragement has been the cornerstone of my perseverance, and their understanding has been a pillar of strength during the challenges of this research endeavor.

In addition, I extend my sincere thanks to my advisor, Dr. Fitsum Assamnew, for his invaluable guidance and expertise, as well as to Addis Ababa University for providing the necessary resources and facilities.

This research journey has been enriched by the blessings of God, the love of my family, and the support of many individuals and institutions. I am truly grateful for the collective contributions that have shaped this work.

Abstract

The importance of yield prediction was understood by several researchers who have proposed different yield prediction models. Prediction results of models depends on data type which could be either linear or nonlinear,spatial or temporal dependent data type, climatic conditions,agronomic practices and soil properties. In addition to the data type if an attention mechanism is added to the model its form of application and of which type, the structure of the model chosen for the work along with the type of suitable optimizer for the model are also other elements on which prediction result of the model depends on. Seeing the gaps that could not be filled by earlier research works in regards to using a fixed attention mechanism and optimizer type, we propose a HYBRID yield predicting model that uses ARIMA prediction and residual as input to the BILSTM model that results in an output where an attention mechanism that is best for the work through a selection process is applied on. ADAM, RMSprop and SGD are the optimizers we have iterated and selected through for this work that best helps the model.Wheat dataset that includes NDVI, weather and yield data is used and its performance was assessed using MSE,MAPE,MAE and MAPE metrics.We made use of ARIMA and BILSTM without attention mechanism as our baseline models for performance comparison.

Two additional model approaches are also discussed in our work one being a Window Sliding approach where BILSTM Model is applied right after the failure point for ARIMA model is known by setting the window size(rolling window) and performing dynamic threshold failure detection using ARIMA validation RMSE metric result as a reference.Any value that is twice as much as the RMSE validation result is taken as a failure point. While the other model approach is a BILSTM model with an attention mechanism added to it. The HYBRID model showed an excellent performance on all metric results by having a lower result in comparison to metrics of ARIMA when it resulted in 98.21 percent on MSE ,a 89.43 percent on MAE, a 86.63 percent on RMSE and a 89.94 percent on the MAPE lower in comparison to its counterpart. In contrary the BILSTM model without attention mechanism which showed lower metric results than the HYBRID when it resulted in 72.54 percent on MSE,a 36.97 percent on MAE, a 31.35 percent on RMSE and a 23.72 percent on MAPE. During performance comparison of the HYBRID model against the Window Sliding model it was found that the HYBRID achieved a lower error value on all the metrics by having a 61.15 percent on MSE, a 37.27 percent on MAPE,a 37.67 percent on RMSE and a 30.41 percent on MAE .In addition we also found the failure point for the ARIMA model is at the middle of the year of 2016 for the Window Sliding model.

Keywords: Normalized Difference Vegetative Index (NDVI),ARIMA-BiLSTM

Table of Contents

List of Figures

List of Tables

List Of Acronyms

BILSTM	Bidirectional Long Short-Term Memory (BiLSTM) network
ARIMA	Auto regressive Integrated Moving Average
MSE	Mean Squared Error
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
ATT-M	Attention mechanism
NAN	Not a Number
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
SGD	Stochastic Gradient Descent
RMSprop	Root Mean Squared Propagation
Adam	Adaptive Moment estimation
n features	Number Of Features
L1 Regularization	Lasso Regression
L2 Regularization	Ridge Regression
ADF	Augmented Dickey Fuller

Chapter 1

Introduction

wheat is among the most important crops grown in Ethiopia, both as a source of food for consumers and as a source of income for farmers. Regarding the area of cultivation, wheat is the fourth most widely grown crop after teff, maize, and sorghum. In terms of the gross value of production, wheat is ranked 4 th or 5th, after teff, enset, and maize and approximately tied with sorghum [1]. NDVI or Normalized Difference Vegetation Index is a remote sensing method that uses the reflectance of light in the visible and near-infrared (NIR) wavelengths to determine the amount and health of vegetation in an area in this work's case the wheat crop's vegetative state. And due to its versatility and ease of extraction, it is a frequently used vegetation index among farmers and researchers [2].It is also widely used in agriculture, forestry, and ecology to monitor the growth and health of vegetation and to identify areas of stress or damage. NDVI values can also be used to map and classify vegetation types, and to detect changes in vegetation cover over time [3].

Machine learning has become one of the most vital decision supporting tools for different areas involved in crop yield prediction starting from helping farmers make informed decisions about what crops to grow and estimating the number of crops that will be produced in a given area to predicting how to sell produced yield a decision that is based on various factors that one must take into consideration. Factors that can affect yield prediction are grouped in three basic categories known as technological, biological and environmental and further divided such as soil type, weather conditions, and crop management practices etc. Ethiopia's agricultural sector plays a critical role in the country's economy. These factors play a major roll resulting in a yield that is amounting to higher or lower at different times based on their change [4].

understanding and utilizing the changes of theses factors for accurate predictions contributes to the wheat yield productivity. Especially for time series behavior having datum having features such as the NDVI, which represents sequences of data points collected over intervals of time, is inherently a rich resource for deriving insightful patterns and trends. Traditional models have been commonly used for time series analysis, however, their limitations such as not being able to be adaptable for any data type, unable to learn from long dependencies when dealing with nonlinear and complex data patterns[5] has led to the need of wanting a prediction model capable of these limitations.

Machine learning models, especially those designed for time series analysis, harness this data to forecast future yields with remarkable accuracy. These models, through sophisticated algorithms, can identify complex, non-linear relationships within the data, which traditional statistical methods might overlook. Techniques such as ARIMA (AutoRegressive Integrated Moving Average), Long Short-Term Memory (LSTM) networks are commonly employed to model and predict agricultural yields. The integration of machine learning with time series data thus holds immense potential in revolutionizing the field of yield prediction, ensuring a more resilient and productive agricultural sector. The factors that majorly affect this proposed work are environmental factors. This paper describes proposed model of machine learning is applied on a time series data contains weather, NDVI data's that are collected through the different growth stages of the wheat crop and yield historical data.

1.1 Statement of the problem

Crop Yield prediction helps farmers and any stack holder in agriculture industry to make an informed decisions.Traditional methods used for yield prediction, which often rely on historical data and simple statistical models, have shown limitations in capturing the complex and dynamic nature of agricultural environments. These limitations are further amplified by the impact of climate change, which introduces additional variability and uncertainty in crop production. Understanding its importance,several researchers proposed different prediction models that use different features [1][2][14][16][17][36]. When coming to research works in areas similar to the models used in the proposed work is one mentioned on [14]. Where authors for the proposed method which is Multi Head ATT-LSTM stated in using Multi head attention type along with ADAM as an optimizer. The selection of an attention mechanism, the structure for prediction model and optimization technique used for the model influence the prediction result obtained by the model. Gaps seen on this research work and others of using fixed attention and optimizer types need to be assessed and filled in order to come up with a model that can do prediction work in a better way. This study will address the following key challenges:

- What things should be considered in order to do dynamic selection of an attention mechanism type that can effectively help the model in its focus
- Evaluating the performance of the hybrid model against traditional prediction methods to demonstrate its superiority in terms of accuracy and reliability.

By addressing these challenges, the research aims to provide a reliable tool for the prediction of wheat yields, which farmers, agronomists, and policy makers can use to make informed decisions, thus contributing to sustainable agricultural practices and improved food security.

1.2 Research Questions

The following research questions are the focus of this research.

RQ1 What are the strengths and limitations of the hybrid ARIMA-BILSTM model in predicting wheat yield in comparison to traditional prediction models?

RQ2 How does the attention mechanism amplify the prediction achieved by the HYBRID model?

RQ3 What other model kinds can discussed to address the performance gap that will be seen if Hybrid model is not effective in its performance?

RQ4 Which environmental features which are highly impacting the Yield?

1.3 Objectives

1.3.1 General Objective

The general objective of our work is to predict wheat crop yield using the hybrid ARIMA -BILSTM model based on different growth stage measured time series NDVI values.

1.3.2 Specific Objectives

The research paper has the following specific objectives:

- The use of the time series NDVI values found in our dataset for the different stages of the wheat and weather data in order to build the Hybrid ARIMA-LSTM model
- Generate prediction and residuals form the ARIMA model.
- Do correlation matrix for understanding feature importance in yield prediction.And do similar relation showing analysis for the non numerical features found in the dataset.
- Generate prediction and residual for the BILSTM model based on its inputs of selected features,de seasonalized yield data,ARIMA prediction and residual and original data.
- Applying an attention mechanism to the BILSTM model's output for focusing the model.
- Evaluating the prediction made by the BILSTM

1.4 Contribution

Our research work aimed to assess the impact of seasonality and trend pattern (periodicity) found in the dataset of weather and NDVI in wheat yield prediction. As a result the main contributions of this work are the following:-

- Assesses the change in optimizers choice done by the different models leading to different outcomes.
- Assesses the impact of not using non numerical data in comparison to converting them and using them as it helps in finding meaningful metric results that actually contribute to the ability of the model
- Assesses also the impact of using a fixed attention mechanism at the output of the hybrid model in comparison to using criteria based selection of suitable attention mechanism.

1.5 Scope and Limitation

In general, the following core points can be summarized on the scope of this thesis:

- Preparing the dataset for the proposed time series yield prediction.
- Examining whether the identified features have an impact on the overall yield prediction.
- Examining the capability of the proposed approach through comparison with existing prediction models.

This research focuses on time series wheat yield prediction and its scope is limited to building a hybrid model that only uses historical yield data along with weather data. Other affecting factors such as soil, water, fertilizers, and pesticides data etc. are not used as part of our dataset. Furthermore, the hybrid model that is proposed is currently not behaving well in comparison to the other models despite many changes to the different parameters found in its structure.The major limitation for our work is the choice of architecture we chose for our model in the fact that applying the attention mechanism in addition to the ARIMA output being used as an input has made the structure complex enough that its performance degraded as a result. The experiments we did highlight this issue due to the other models not using any of the ARIMA output has resulted them in having a better performance when compared to our proposed model.

The other limitation we believe is related to the period when the BILSTM model starts to do its prediction, this is further explained from the experiment we did for the Window Sliding model where it starts after detecting the inability of the ARIMA model in precise manner whereas in our proposed model the BILSTM starts its prediction after all the necessary inputs are forwarded for it this we believe also created a lowering in performance for the proposed model.

1.6 Organization of the study

The rest of this document is organized as follows. The second chapter discusses definitions,types,theoretical backgrounds related to time series data and its relation to yield prediction. In addition to that previous research works related to yield prediction and state of the art approaches used are also discussed. Our proposed approach for wheat yield prediction is elaborated in Chapter three. The experimental setups, procedures, evaluation metrics, results, and discussion are discussed in Chapter four. The final chapter discusses the conclusion and future research direction on hybrid model wheat yield prediction.

Chapter 2

Literature Review and Theoretical Background

This chapter discusses the theoretical background of this study and integrates concepts from time series analysis, remote sensing, weather data utilization, and hybrid modeling approaches to develop an advanced predictive model for wheat yield. The chapter also discusses the different works that were done on topics related in this research work.

2.1 Time Series Analysis

Time-series analysis involves the study of data points collected or recorded at specific time intervals. The purpose of this project is to identify patterns, trends, and seasonal variations to make predictions about future data points. It differs from other analysis types by its ability to identify the characteristics of the district that lead to a specific pattern in the data series points. The initial step in time-series analysis is the process of determining if the series is stationary or not. The existence of trend and seasonality in the time series results in making the mean, autocorrelation, and variance constant throughout time, as a result it is defined as stationary .However, if these properties don't show this sign then the series is termed as non-stationery[6].

2.1.1 Trend

The trend component in a time series refers to the long-term progression of the data, illustrating a consistent upward or downward movement over time. This component captures the underlying direction of the data, excluding short-term fluctuations and seasonal effects. It has the following characteristics:-

- Long-Term Movement: Trends depict the overall direction, whether increasing, decreasing, or stagnant, and are influenced by underlying factors like technological advancements, economic conditions, and demographic shifts.
- Non-Periodic: Unlike seasonal patterns, trends do not repeat at regular intervals.
- Deterministic or Stochastic: Trends can be deterministic (predictable and steady) or stochastic (random and evolving).

2.1.2 Relevance to Wheat Yield Prediction

- Understanding the trend component in wheat yield data helps identify long-term agricultural productivity changes due to factors like advancements in farming practices, changes in crop varieties, or persistent climate changes.
- Trends in NDVI values can indicate long-term changes in vegetation health and biomass, influenced by sustained environmental conditions or agronomic interventions. The following tests will be conducted to check whether or not a time series is stationary;
- Plotting Rolling Statistics: this testing mechanism Plots the moving average or difference and checks whether it depends on time (time varies). It is more like a visual representation.
- Dickey-Fuller Test is a statistical test used to determine the stationarity of time series, unlike the first one. This test assumes the null hypothesis is not stationary to the time series. The test-statistic and some critical values for various confidence levels are the results. When the test-statistical falls below the critical value, the null hypothesis is rejected, the series is stationary.

2.1.3 Seasonality:

The seasonality component refers to periodic fluctuations in the time series data that occur at regular intervals, typically within a year. Seasonal patterns are influenced by recurring events or cycles, such as weather conditions, holidays, or agricultural cycles. It has the following characteristics:-

- Regular Intervals: Seasonal effects repeat at fixed periods, such as monthly, quarterly, or annually.

- Predictable Patterns: The magnitude and direction of seasonal variations are generally predictable, making them crucial for accurate forecasting.
- Additive or Multiplicative: Seasonal components can be additive (constant size over time) or multiplicative (varying size relative to the level of the series).

2.1.4 Relevance to Wheat Yield Prediction:

- In agriculture, seasonality is prominently seen due to the natural growth cycles of crops, planting and harvesting seasons, and climate variations throughout the year.
- Seasonal patterns in NDVI values reflect the annual growth stages of wheat, from sowing to maturity, providing critical insights into crop health at different times of the year.
- Weather data also exhibits seasonality, with periodic changes in temperature, precipitation, and other climatic factors significantly impacting crop yield.
- The primary methods used in time series analysis include ARIMA and LSTM.

2.1.5 Strategies for eliminating trend and seasonality components:

- Differencing: is the most common strategy to reduce non-stationary features by calculating the difference between a current observation and the previous one. The sequence of differences can be modified to optimize trend and seasonal reduction.
- Decomposing: is the tool for all kinds of time series analysis, especially seasonal adjustment. It is intended to build a component that could be employed by adding or multiplying the original data in an observation series. A distinct sort of behavior characterizes each of the features.

2.2 Remote Sensing and NDVI

Remote sensing technology enables the collection of data about the Earth's surface without physical contact. The Normalized Difference Vegetation Index (NDVI) is a widely used remote sensing index that measures vegetation health and biomass:

- NDVI Calculation: NDVI is calculated using the near-infrared (NIR) and red (R) wavelengths of light reflected by vegetation. The formula is:

$$NDVI = \frac{NIR - R}{NIR + R} \tag{2.1}$$

  where Higher NDVI values indicate healthier and more vigorous vegetation.

- Applications in Agriculture : NDVI is used to monitor crop growth, assess plant health, and estimate yield potential. It provides valuable insights into vegetation dynamics and can be used to predict crop yields.

2.3 Weather Data in Agriculture

Weather conditions play a critical role in agricultural productivity. Factors such as temperature, precipitation, humidity, and solar radiation directly impact crop growth and development. Incorporating weather data into yield prediction models helps:

- **Understanding Climate Impact:** Weather data provides context for variations in crop performance, allowing for more accurate yield predictions.
- **Enhancing Model Accuracy:** By integrating weather variables, models can better account for environmental influences on crop yield.

2.4 Feature selection Algorithm

Feature selection is the process of selecting a subset of relevant features from the original feature set.Due to the challenges arising from excessive, redundant and irrelevant dimensions and the presence of noise, it has become an indispensable component in building robust models. In fact,feature selection has been playing an important role in many applications since it can speed up the learning process they allow the model to run as quickly as possible with the least amount of time complexity while still providing the best output with the best predicted values, improve the model generalization capability by reducing over fit,alleviate the effect of the curse of dimensionality, and simplify the model for easier interpretation [4].

In terms of the availability of label information, feature selection algorithms can be roughly classified as supervised,unsupervised, and semi-supervised methods. In the super-vised case, the labels of training data are used to select discriminative features to distinguish samples from different classes. In the unsupervised case, feature selection becomes more challenging due to the lack of label information. As a compensation of label deletion, a variety of criteria have been proposed to define feature relevance. In semi-supervised case, both labeled and unlabeled data are utilized to choose the optimal feature subset. According to, in terms of algorithms, feature selection methods can be categorized into four groups, i.e. , similarity based, information theory based, statistics based and sparse learning based methods. Similarity based methods evaluate the importance of features by its capacity to preserve data similarity. Information-theory-based methods exploit different heuristic filter criteria to measure the importance of features. Statistical methods evaluate the relevance of features by designing a variety of statistical measures. Sparse learning-based methods aim to minimize the fitting errors along with some sparse regularization terms. The sparse regularizer forces some feature coefficients to be small or exactly zero, and then the corresponding features can be simply eliminated [7].

2.4.1 Auto Regressive Integrated Moving Average (ARIMA)

ARIMA models are popular for their ability to handle various types of time series data. They work by combining three components: auto regression (AR), differencing (I), and moving average (MA). ARIMA is effective for modeling linear relationships in time series data.

- Trend Component: ARIMA models effectively capture linear trends by differencing the data to make it stationary, removing long-term dependencies.
- Seasonal Component: Seasonal ARIMA (SARIMA) models extend ARIMA to account for seasonal variations by incorporating seasonal differencing and seasonal autoregressive and moving average terms.

2.5 Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) capable of learning long-term dependencies. It is particularly useful for sequential data where previous information influences future values. LSTM networks are effective in capturing non-linear patterns in time series data.

- Trend Component:LSTM networks are capable of learning and predicting complex, non-linear trends through their ability to retain information over long sequences.
- Seasonal Component: LSTM models can capture intricate seasonal patterns by learning periodic dependencies in the time series data, adapting to the varying seasonal effects.

Time series prediction is the prediction of the future value over a period of time it involves the development of models based on past time values and using them to get prediction of feature value. Sequential nature of the data is taken into consideration where each observation is dependent on previous observations. It has the following three characteristics:-

1. Temporal ordering: It is ordered chronologically where each observation or event occurs after the previous one.
2. Time dependency: Time series has a sequential relationship where a value at current time depends on similar values that occurred previously.
3. Irregular sampling: This is when an irregular or uneven time intervals exist between observations.

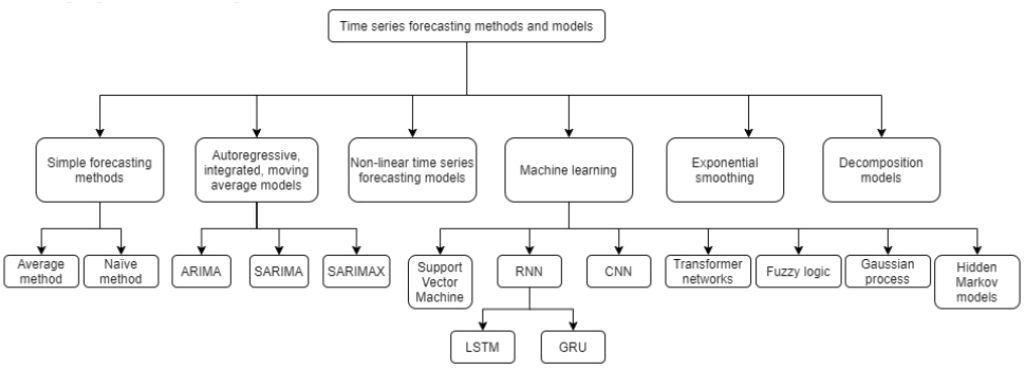

Figure 2.1: Classification of time series forecasting methods and models

Based on the technical background, characteristics and classification of the models shown above on Figure 2.1 for the classification of time series forecasting methods and models there have been many research works done in the past, we have listed some of them below.

2.6 Yield prediction using Autoregressive, integrated, moving average models

An autoregressive integrated moving average refers to a statistical analysis model utilizing time series data to understand the data set better or project future trends. As seen from the above figure one of the category of these models is ARIMA model which combines three components auto regression (AR), differencing (I), and moving averages (MA)—to describe and predict future values based on past observations. Identifying time series stationary, identifying the order of the AR and MA, determining of p,d,q , performing diagnostic checks and finally predicting are the major steps usually involved in building these models.

Fan, Chao and Cao, Pei-Ge and Yang, Tie-Jun and Fu, Hong-Liang [8]. proposed that arima model performed better in comparrision to the Grey model and second exponential smoothing method after analyzing yield data from 1980- 2012. The AIC was used for selecting from the three ordered arima models of ARIMA(5,1,1), ARIMA(3,1,5) and ARIMA(5,1,5) where upon testing before the comparison with the other two models which were Grey Forecasting model GM(1,1) and the second exponential smoothing after this comparison for accuracy the authors concluded that ARIMA(5,1,5) order was chosen for achieving the highest accuracy. The papers failure in not showing additional machine learning models that are far much much proved better for comparison and actually for seeing if actually proposed and selected model is a better fit for the data.

Sowmya and Prasad[9] proposed a methodology of combining ensemble technique with ARIMA model for crop yield prediction that answers the challenges of data sparsity and provides prediction in fine detail in both state and district. The ensemble technique uses Random forest regress or for yield prediction. Proposed methodology contributes to precise agriculture and sustainable planning in applied region.Factors such as data quality, limited variables, and the dynamic nature of agricultural systems were challenges that were faced by the authors but each factor's effect on the selected model was not mentioned in the entire work which is a major limitation on the work.But the reflection for other researchers on future work is one that can be taken as a positive side of the paper.

Kannan, S and Karuppasamy, KM [10] authors concluded that ARIMA model was a robust tool for agriculture prediction after achieving high degree of accuracy and reliability. These results were achieved by basing on the results that were found through the application of metric such as mean squared error(MSE) and mean Absolute error (MAE).The prediction was done in 4 different states with ARIMA(0,1,1), ARIMA(0,1,1), ARIMA(0,1,2) and ARIMA(0,1,1) being the different models for reaching the above conclusion. Neither of the ARIMA models were compared with other time series models for actually accepting the conclusion that was reached. And unlike the [2] both this and the [1] did not reflect future angels of exploration that might help other researchers working on similar topics

Rathod, Santosha and Singh, KN and Arya, Prawin and Ray, Mrinmoy and Mukherjee, Anirban and Sinha, Kanchan and Kumar, Prakash and Shekhawat, Ravindra Singh [11] The integration of arima and Genetic algorithm (GA) is a powerful combination for predicting the maize yield forecasting. In addition, this combination outperformed traditional methods in terms of prediction accuracy and reliability. The Arima model is used for the temporal patterns of the yield data while the GA was applied to optimize the parameters of the Arima model so as to increase the natural selection process. Similar to [1] and [3] this paper also fails to validate the conclusion reached about the model through comparison with other models.

2.7 Yield prediction using Hybrid Machine learning models with addition of an Attention mechanism

The following mentioned below are some of the research works that is believed to be done in areas that are related to the current proposed work.

Cho, Wanhyun and Kim, Sangkyuoon and Na, Myunghwan and Na, Inseop[12], concluded that combining LSTM networks with attention mechanisms and ARMA models can significantly enhance the accuracy of tomato yield predictions. The hybrid model was constructed in a manner where an encoding attention-based LSTM is used for identifying environmental factors that affect the time series where as the ARMA model as a statistical time series analysis model to improve the difference between the actual yields and the predicted yields given by the attention-based LSTM model earlier. The hybrid model was used for tomato yield prediction and showed it performed well in comparison to existing models. Additionally through different sets of experiments it was found that internal temperature, internal humidity,and CO2 level were the environmental factors that highly affected the tomato yield. One major limitation that is seen on this research work is the fact that it does not take in to consideration temporal weighting for its variable despite of them all being a time series based ones.

Cui, Ligang and Chen, Yingcong and Deng, Jie and Han, Zhiyuan. [13] the hybrid model proposed on this paper is a combination of enhanced bidirectional LSTM and self-attention mechanism which outperformed seasonal autoregressive integrated moving average, support vector machine, random forest, and LSTM models that were used for comparison. This conclusion was reached through the different set of experiments. The Bi-LSTM network captures temporal dependencies from both past and future data points, enhancing the model's ability to understand the sequence while the attention mechanism helps the model focus on the most relevant parts of the input data, improving prediction accuracy. The authors concluded that the proposed model showed robust generalization capabilities in uni variate time series demand forecasting and a powerful tool for demand forecasting, offering valuable insights for supply chain decisions and operational planning. The fact that the authors used a diversified dataset for their work is one major point that is mostly not seen on the other papers but the metric evaluation mechanism only being two which are RMSE and MAPE and in addition to not giving a detailed explanation on their choice posses concern on the conclusions that are reached which is a week point that is taken from this work.

Abbasimehr, Hossein and Paki, Reza. [14] paper addresses the challenge of accurate time series forecasting, which is crucial in many sectors. Traditional forecasting methods often struggle with non-linear patterns and complexities that is found in a real-world data. To overcome these, the authors proposed a hybrid model combining Long Short-Term Memory (LSTM) networks and multi-head attention mechanisms. The LSTM component captures long-term dependencies in the data, while the attention mechanism helps the model focus on relevant parts of the input sequence, improving its predictive accuracy. The proposed model is evaluated on 16 different datasets and compared with standard forecasting techniques such as exponential smoothing (ETS), ARIMA, multilayer perceptron (MLP), as well as other individual models based on LSTM and multi-head attention are used in addition to those proposed by (Babu and Reddy 2014; Panigrahi and Behera 2017; Zhang 2003). The results of experiments on public time series data demonstrate that the proposed hybrid model outperforms other utilized in terms of symmetric mean absolute percentage error (SMAPE). Moreover, the results of comparisons suggest that the method developed based on multi-head is the second best model. The results demonstrate that the hybrid LSTM-attention model outperforms all benchmark methods in terms of Symmetric Mean Absolute Percentage Error (SMAPE) and achieves the best average rank among the tested methods proving the importance of an attention mechanism. A limitation that can be taken out for this work is the fact that the authors usage of Bayesian optimization technique to reduce the experimentation of their work makes the achieved results incomplete since without using this technique the conclusions possibly might change which is similar thing shown on paper [18] with usage of Google Colab.

Liang, Luocheng. [15] proposed a hybrid model of ARIMA with Attention-based CNN-LSTM and XGBoost that focuses on stock price prediction it effectively combines the strengths of each component leading to improved prediction accuracy and robustness. The ARIMA model was used for initial data preprocessing to handle linear trends and seasonality while the Attention-based CNN extracts deep features from the preprocessed data, focusing on relevant parts of the input sequence while the LSTM deals with the non linear trends and seasonality part. The XGBoost is employed for fine-tuning the predictions. The model is evaluated on stock data from five stock corporations in the US market, and comparisons were made on how this model performs on the five different data types obtained from the corporations. Based on the results obtained from MAE and MSE show that it performs a lot better on predicting stocks with larger price fluctuations in comparison to those having lower price fluctuations. The limitation of this paper is the fact that since the author only used a limited dataset the proposed models actual ability could not be concluded in regards to its ability in prediction. In addition to that like [1]and [4] it does not mention comparing experiments that are done with other predicting models so as to truly accept the model as the best one.

2.8 Yield prediction using Hybrid models without attention mechanism

A high performance was reached by a hybrid model of CNN-DNN when compared along with XGBoost, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), CNN-XGBoost, CNN-Recurrent Neural Networks (RNN), and CNN-Long Short-Term Memory (LSTM) models built for yield prediction by Oikonomidis, Alexandros and Catal, Cagatay and Kassahun, Ayalew[16] The experiments that were used for these conclusions were performed on a soyabean dataset with 395 features including weather and soil parameters and 25,345 samples. The hybrid models performance was achieved with the results of the metric tests of RMSE.MSE,MAE and R2. The limitation that is faced by the authors was the limited experimentation environment that was used due to the fact that the authors used the free package usage of the Google Colab which bounded them not to further observe the models abilities.

The research paper done by Agarwal, Sonal and Tarar, Sandhya[17] showed a hybrid model of machine learning (SVM algorithm) and deep learning (LSTM, RNN) techniques which predicts the best suitable crop that needs to be grown on a certain area based on the dataset that contains multiple crop types such as wheat, rice, maize, millets, green gram, pea, pigeon pea, sugarcane and many more. Additionally a few prediction parameters like, pH value, temperature, rainfall, relative humidity, and area is also found in the dataset for each crop type. The authors concluded that Machine Learning algorithm along with the Deep Learning algorithms play a vital role in predicting the yield with better accuracy of 97 percent better when compared to the other hybrid model they used on their experiments which was the Artificial Neural Network (ANN), and Random Forest algorithms,which gave an accuracy of 93 percent. Even though the authors tried to explore the perfect condition for the best yield predict which is a good foundation for any future works that base it. like [14],[4],[1] the fact they only considered a single model for their comparison limits the models true abilities especially if it were to be compared with more advanced techniques.

2.9 Yield prediction using Non hybrid Machine learning models with addition of an Attention mechanism

Wu, Pan and Huang, Zilin and Pian, Yuzhuang and Xu, Lunhui and Li, Jinlong and Chen, Kaixun. [18] proposed an attention-based long short memory(ATT-LSTM) model for predicting the short term traffic speed. It was basing on the generalization that urban traffic speed prediction is a challenge due to its temporal, spatial correlations and complex nonlinearity they concluded that the ATT-LSTM model outperforms other deep learning algorithms such as RNN and CNN in computational efficiency and prediction accuracy. The combination improves the short-term traffic speed prediction by becoming a valuable tool for intelligent transportation systems. The attention helped the LSTM in making it give a better attention by assigning weights to important traffic time. Two limitations that are seen on these research work is the fact that the dataset is limited in the number of features and the dataset only contains a 30-day traffic speed dataset and not previous historical years which might change the outcome reached for the proposed model.

The LSTM model enhanced with an attention mechanism that is added for capturing the temporal dependencies was Proposed by Geddlehally Renukaradya, Nandini and Rao, Kishore Gopala and Jayachandra, Anand Babu [19] for yield prediction model. The model achieved high accuracy when checked by MAE,MSE,RMSE in addition to accuracy and R2. For the development of the work the authors used weather data, soil properties and other variables that went through techniques such as correlation-based feature selection algorithm, variance inflation factor for selection relevant features and removing of any Multicollinearity. The authors used three different crop types which covers the biological factor aspect of yield prediction. The major limitation of the paper is the fact that the authors failed to give a detailed description on how the authors used the Convolutional Neural Network (CNN),Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Generative Adversarial Network (GAN) models which were used for performance comparison.

Zhou, Kun and Wang, Wen Yong and Hu, Teng and Wu, Chen Huang[20] Their model tries to show the effectiveness of two prominent time series forecasting models: the traditional statistical ARIMA (AutoRegressive Integrated Moving Average) model and the deep learning-based LSTM (Long Short-Term Memory) with an Attention Mechanism. The aim is to compare their performance in predicting complex time series data. The paper showed the advantages of combining LSTM networks with attention mechanisms for time series forecasting, showcasing their effectiveness over traditional ARIMA models, especially for complex and non-linear time series data. These advantages were discovered using the different kinds of optimizers like adadelta, adagrad, adam, rmsprop, sgd and adam optimizer . In addition the ARIMA model performed better that the LSTM with attention for small-scale of TSF work by keeping network structure, traffic characteristics, accuracy, complexity and other factors in check.The limitation of the paper is the authors only doing experiment in changing the different optimizer while keeping the many influencing factors such as learning rate, epoches etc that could have changed the conclusion that was reached by the authors.

A deep learning-based model that combines Long Short-Term Memory (LSTM) networks with an attention mechanism for addressing the challenge of accurately predicting grain yield in China was proposed by Liu, Fan and Jiang, Xiangtao and Wu, Zhenyu [21].The model uses MODIS remote sensing image data from 2010 to 2020, incorporating vegetation indices and temperature data to form a composite dataset while Convolutional Block Attention Module (CBAM) is employed to extract spatial information from different remote sensing bands, improving the model's efficiency . The work showed that combining LSTM networks with attention mechanisms can significantly improve grain yield prediction by leveraging multi-source satellite imagery through the results obtained from RMSE and R2 Score metric tests.Even though the authors used 28 provinces crop data that is having spatial and climatic data of which the dataset duration being from 2018-2021, the authors chose all their features in addition to the mentioned ones by taking other researchers work and failing to do any correlation analysis so as to asses which attributes contribute and which not for the yield prediction work making it a limitation that is found on their work.

Another paper that made use of adding an attention mechanism for better yield prediction is one proposed by Xiang, Wei and Long, Long and Liu, Zichen and Dai, Feng and Zhang, Yucheng and Li, Hu and Cheng, Lin[22] which aimed at making a model for yield prediction that is based on an attention mechanism and that utilizes weather and soil data. The attention mechanism here has two modules one for extracting temporal features of crop growth in different regions and the other being the feature attention module that extracts environmental factor features in different regions. Together, the two modules extract spatial differences in different regions. The multi-layer perceptron model is what is used for soybean yield prediction after using the two modules output as an input. The authors concluded that after achieving lower root mean square error (RMSE) and higher correlation coefficients compared to CNN, LSTM and Random forest. Two limitations that can be taken out from this work is the model actual ability to adapt to significant spatial differences in crop growth environments that is found across different regions and concluding the model's ability to predict will work without doing any further experimentation for any other crop type.

Shook, Johnathon and Gangopadhyay, Tryambak and Wu, Linjiang and Ganapathysubramanian, Baskar and Sarkar, Soumik and Singh, Asheesh K[23]proposed a deep learning-based model for predicting crop yield by integrating genotype and weather variables. The model employs a Long Short-Term Memory (LSTM) network combined with a temporal attention mechanism to capture temporal dependencies and important time-windows in the growing season. The predicting ability of the proposed model was compared along with Support Vector Regression with Radial Basis Function kernel (SVR-RBF), least absolute shrinkage selection operator (LASSO) regression , data-driven USDA model and it showed great result based on the results obtained from MAE,RMSE and R2 Score. A limitation that is seen on this work is the lower number of training dataset used for this work which will lead to over fitting despite comparison conclusion stated earlier.

The research paper done by Tian, Huiren and Wang, Pengxin and Tansey, Kevin and Han, Dong and Zhang, Jingqi and Zhang, Shuyu and Li, Hongmei [24] which had the objective of creating a deep learning model that can do winter wheat yield estimation based on weather data , Vegetation Tem perature Condition Index (VTCI) and Leaf Area Index (LAI) at the main growth stages of winter wheat. The proposed model is an attention mechanism based LSTM and its effectiveness in achieving the set objective was shown through a comparison that is made against an LSTM model. And based on results that are obtained from the set of tests of R2, MAPE,RMSE,NRMSE it was concluded that the proposed model was seen to be predicting the yield in a much better way than the LSTM model. In addition to the above conclusion the authors also noticed that the proposed model adapted very well to different kinds of sites having different conditions of either being rain fed or irrigated one. They also were able to identify which of the remote sensed variables contributed to the yield more by stating LAI at the heading-filling stage and the milk stage as well as VTCI at the jointing stage contributed more than other input feature variables. A positive part that can be taken out from this work is the fact that the authors actually did further investigation of which spatial parameters affect the wheat crop and decided based on the effects they had on the critical stages of the crop. But when it came to weather data the fact that they only took temperature and perceptiation with out any further analysis like the spital data can be taken out as the weakness of the paper in similar ways to paper [15],[20] and [28]

2.10 Yield prediction using Combination of machine learning models

The model proposed by Osibo, Benjamin Kwapong and Ma, Tinghuai and Wahab, Mohamed Magdy Abdel and Jia, Li and Wenzheng, Ye and Bediako-Kyeremeh, Bright and Osei-Appiah, Stephen [6] built for yield prediction by using historical wheat yield data, climatic conditions, soil data and moderate resolution imaging spectro radiometer(MODIS) data. The model was a combination of LSTM network with Extreme Gradient Boosting (XGB) and used SHAP (shapely additive explanations) tool for prediction of the yield. The prediction achieved by this model showed a much better result than other models that were built solely for this comparison like LSTM, Light Gradient Boosting Machine(LGBMR) and Deep neural network.Unlike paper[6] the authors comparison of the proposed model with more than one model that could do the same work is taken as a positive side of the paper whereas like paper [19]and [34] the dataset only comprising of that area brings questionable remarks on the performance of the model when exposed to a dataset that is outside of Egypt.

An ensemble model of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) for cocoa yield prediction by Olofintuyi, Sunday Samuel and Olajubu, Emmanuel Ajayi and Olanike, Deji [25] concluded that the ensemble model was effective in comparison to other machine learning models such as NB,MLP, SARIMA and LSTM. In addition to these comparisons other models incorporating LSTM were also used for demonstrating its effectiveness these combination models were RNN-LSTM, CNN-LSTM. In this work the CNN is used for processing the climate dataset that is used while the RNN with the RNN is used for the cocoa yield dataset and making the ensemble model predicts the yield by addressing the spatial and temporal patterns of the data. Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) were the measuring metrics for the above stated conclusion that was used by the authors. Like paper [34] proposed model's ability to cope with different climatic conditions found in other areas outside of southwest Nigeria is in question since work is only limited to this section of the country since like paper [18] cost for climate data collection has become high for authors work making it a con side of the paper.

LSTM-NN forecasting model was proposed by Haider, Sajjad Ali and Naqvi, Syed Rameez and Akram, Tallha and Umar, Gulfam Ahmad and Shahzad, Aamir and Sial, Muhammad Rafiq and Khaliq, Shoaib and Kamran, Muhammad [26]. For wheat production the authors used data pre-processing smoothing mechanism of Robust-LOWESS in addition to it. The model was also used for showcasing Pakistan's major problem of the production to consumption ratio for wheat. The model showed it could do the intended target of production forecasting through the results obtained from the R-value ,MAE,RMSE metric tests.The model was compared with ARIMA and RNN forecasting models and showed great output. Despite the fact that parameter tuning plays an important roll for works such as this the authors mentioned almost to nothing on the different experiments they took for selecting the parameters needed for the work which shares a limitation with paper[15].

PATIL, PRITESH and ATHAVALE, PRANAV and BOTHARA, MANAS and TAMBOLKAR, SIDDHI and MORE, ADITYA[27]proposed a model where one model for Crop prediction which is done by classification model while yield prediction uses regression models to learn from the data. Among the many ML classifiers like Logistic Regression, Naïve Bayes, Random Forest, KNN were used and compared with the performance metrics and the model which achieved best accuracy was selected for crop prediction while for Yield prediction, regresses like Linear Regression, Random Forests and Decision Trees were used and based on the metric tests of MAE, Median Absolute Error and R2 Score Random Forest Regression gave best results while for crop prediction, Naïve Bayes classifier gives most accurate results. A point that is taken as a limitation is not using real time data which made the authors constrained and unable to compare with real time data using models that are more robust so as to actually test its ability.

2.11 Yield prediction using individual machine learning models with NDVI data

A research paper aiming in developing a yield predicting model that leverages weather parameters and NDVI data to accurately forecast crop yield was proposed by Galphade, Manisha and More, Nilkamal and Wagh, Abhishek and Nikam, VB [28]. For their work weather and soil parameters, normalized difference vegetation index (NDVI) is used and is subdivided into three different sections which are (I) Prediction of the weather parameters, (II) prediction of NDVI using weather parameters as input, (III) yield prediction using stage I and II as input. Based on the results they got from their set of experiments for the for the prediction of the first and second subsections of the model KNN model achieved better result in terms of accuracy whereas in the prediction for the third subsection of the model Decision Tree was found to be accurate in tits results than other models. For reaching these conclusions the authors made use of RMSE and R2 metric tests. And like paper [1],[4]and [14] failure to do valid comparison with models that can achieve a more robust result is the limitation that can be taken out for this work.

Another paper that made use of NDVI time series data for yield prediction is that written by Barriguinha, André and Jardim, Bruno and de Castro Neto, Miguel and Gil, Artur [29]the work revolved around creating a Using a satellite-based time-series of Normalized Difference Vegetation Index (NDVI) calculated from Sentinel 2 images and climate data acquired by local automatic weather stations, and finally using a system for vineyard yield prediction based on a Long Short-Term Memory (LSTM) neural network. In addition to the NDVI data data's of Temperature, Relative Humidity, Precipitation, and Wind Intensity were also considered for the work. Two particular growth stages were considered for the work namely FLO (flowering) and VER (veraison). Upon the different tests done namely MSE and MAE it was noticed that VER period or growth stage was chosen to show the vineyard prediction in a much better way. The authors concluded that using NDVI alone is insufficient for a robust and accurate model developed with their proposed methodology and climatic variables, using satellite data and meteorological variables, environmental conditions and management strategies should all be considered for achieving result that is much better than found from their experiments.The fact that the dataset was small due to low yield data in addition to weather stations where data is collected being small are major limitations that are found on the work contributing to not having a more refined result than what is found.

2.12 Yield prediction through comparison by individual machine learning models

Comparison between ARIMA and LSTM was shown on Albeladi, Khulood and Zafar, Bassam and Mueen, Ahmed [30] research paper once again and LSTM's ability to dealing with complex and non-linear parts of the time series data has made it a better model. Even though the dataset that is used for this work is Mulkia Gulf Real Estate from Saudi Exchange datasets and not related to crop prediction it is one of the many researches that are done that show time series forecasting or prediction using these two models. Metric tests such as R2 Score, MSE,RMSE,MAE,MAPE,MDAPE were used in addition to avg-loss and val-loss graphs to evaluate models like LSTM and for reaching the above conclusions. Like [1],[4],[14] and [36] this paper's consideration of only ARIMA for comparison with the LSTM is not a valid comparison since it is biased that LSTM performs better usually.

Pravallika, K and Karuna, G and Anuradha, K and Srilakshmi, V[31] proposed a DNN (Deep Neural Network) model for yield prediction. The model is designed to learn patterns that are termed to complicated for other models. To make archive this features that highly influence yield of crops are identified so that the model gives its focus to these features than others. This models ability for yield prediction is shown through the tests done such as the MAE, RMSE and R2 Score. The authors work made use of historical yield data in addition to soil conditions, weather and different agriculture parameter's. Even though the length the authors took to describe the different models they used for the performance comparison is commendable, the fact that they have not described in great depth their own model is taken as a weak point on their work since it leads to confusion on what the authors are proposing on the work.

Another research work that compares ARIMA's prediction ability in comparison to LSTM's is the work of Abdoli, Ghahreman [32] done for showing the models ability in dealing with time series data that has fluctuations. From the different set of experiments that were performed the authors of the work concluded even though both models experience a reduction in forecast accuracy over the long term LSTM models are more effective than ARIMA models for time-series data with permanent fluctuations, providing more accurate and reliable forecasts. Tehran stock price was what was used as a dataset by the authors while using metric tests for generating results for the above mentioned conclusion. The limitation that can be taken out from this paper is despite the fact that the paper is based on time series data how the data is used for the prediction is not shown in detail creating a black box image for any researcher referencing this work.

The research paper done by etiner, Halit and Kara, Burhan [33] aimed on a comparison by suggesting a Recurrent Neural Network (RNN) model based algorithms especially Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) methods for wheat yield forecasting so as to use its ability to handle sequential data and capture temporal dependencies. The models ability to forecast in an effective manner was evaluated using metrics such as Mean Squared Error (MSE), MAE and MAPE, R2 score and Root Mean Squared Error (RMSE). Based on these results LSTM model was found to be better than the GRU model despite the LSTM obtaining a larger time frame for training in comparison to the GRU. From these metrics it was concluded that RNN based models are effective for wheat yield forecasting, providing a robust tool for agricultural planning and decision-making. The dataset being based on the Konya province of turkey could be a possible limitation of the work since due to geographical change the results obtained for the proposed model might not be the same limiting the use of the research to only that area.

Mo, Haoming and Zhang, Ying and Liu, Yifei and Zheng, Yanzi[34] developed LSTM model showed a much better performance in predicting the yield of rice when compared to traditional method of support vector regression model which by itself is a good yield predictor. In addition the developed model was also compared with standard recurrent neural network (RNN) in which the developed model performed well in solving the problem of gradient dispersion and loss of gradient which is experienced mainly by RNN. In light of this the authors of the work concluded that the developed model's abilty to predict the yield achieved high accuracy when historical yield data was used in addition to metro logical data. A limitation that is observed from this paper and that is shared by paper [1],[4],[14],[36] and [21] is the fact that the authors could have included other models that also deal with gradient loss like the proposed model so as to know exactly how much the proposed model's performance would be.

Having the objective of comparing and evaluating different Machine Learning and Deep Learning models for crop yield prediction for identifying which is the most effective techniques for various crops and conditions were presented by Jhajharia, Kavita and Mathur, Pratistha and Jain, Sanchit and Nijhawan, Sukriti [35].Among the traditional machine learning algorithms that were under observation were Random Forest, Support Vector Machine (SVM), and Lasso Regression while Deep Learning models like Gradient Descent and Long Short-Term Memory (LSTM) were models that were used for the work. The work was based on dataset of seven crop types namely Rapeseed and Mustard, Wheat, Barley, Bajra, Jowar, Onion and Maize. In conclusion the ML models performed better in comparison to the DL models in terms of accuracy due to the fact that these models predicted the yield on the available limited dataset especially Random Forest algorithm was found to be the most effective model for all the crop type datasets that were used for the work. Like other works also metric tests such as RMSE,MAE,R2 Score were also used for reaching the choosing of the right models.Similar to paper [25] the limitation that can be taken out from this work is due to the fact that the dataset is limited it might have an imbalance in terms of crop types, environmental conditions, or geographical locations, leading to biased predictions that favor more frequent categories or locations and conclusion like the one that is stated.

Another paper that also dealt with multiple crop types data type for their work was the one done by Cedric, Lontsi Saadio and Adoni, Wilfried Yves Hamilton and Aworka, Rubby and Zoueu, Jérémie Thouakesseh and Mutombo, Franck Kalala and Krichen, Moez and Kimpolo, Charles Lebon Mberi[36] which had an objective of building machine learning model that can predict the yield of six crop types namely rice, maize, cassava, seed cotton, yams, and bananas. For these work in addition to agricultural yields climatic data, weather data and chemical data were also used. The models used for this work were decision tree, multivariate logistic regression, and k-nearest neighbor models. The authors concluded that from the models the decision tree model performs well with a coefficient of determination (R2) and MAE results obtained when compared with the K-Nearest Neighbor model and logistic regression. In addition the prediction results of the decision tree model and the K-Nearest Neighbor model are correlated to the expected result, showing the effectiveness of the models. The fact that the authors actually included six African countries in their data set is by itself a great contribution especially being able to propose a model that is robust enough to handle the different climate situations of each country. But the fact that they did not describe in great detail the choice of parameters for their dataset among the many factors that affect crop prediction which is the weak side of the work like papers [15]and [20].

In addition to choosing a predictive model for yield prediction , amount of data for the general work and dataset portioning was explored by Morales, Alejandro and Villalobos, Francisco J [37].Their work included having a Synthetic datasets from biophysical crop models (OilcropSun and Ceres-Wheat from DSSAT) were used, simulating sunflower and wheat data from 2001 to 2020 in five areas of Spain. Regularized linear models, random forest, artificial neural networks were the algorithms used for the work by keeping in mind seasonal, management and soil of the experiment area. The authors applied chronological data splitting method for dataset partitioning. It also consisted the effect of extrapolation and interpolation use of data and its outcome on the overall yield prediction by the stated models. From the results obtained from the RMSE and R2 they concluded that for small tabular dataset using random forest was much better than ANN and achieves a good result that is only achieved by ANN models if there is only a large amount of data. The regularization applied with the linear models is not effective while dropout layer applied for the ANN showed great change in the models output.Random partitioning was suggested as not the correct data partitioning method where interpolation is used for any features found in the dataset since it lead to underestimating model errors, as compared to time-dependent partition. Despite the chronological data splitting usage the fact that the authors did not include real time seasonal weather forecasting data can be taken as the limitation of the work since it leaves a doubt for any researcher on what the outcome might have been when trying to reference this work.

Research paper that was written by Reddy, D Jayanarayana and Kumar, M Rudra [38] aimed at how prediction of yield of a crop can be further improved by exploring models that result crop yield estimation based on the weather, crop disease, classification of crops based on the growing phase etc. The authors showed several existing models that consider elements such as temperature, weather condition models such as Neural networks, random forests, KNN regression and a variety of ML techniques such as CNN, LSTM, DNN algorithms. The result showed that crop yield estimation would be improved by further researchers by combining of ML with the agricultural domain field for improving the advancement in crop prediction in addition to improvement of feature selection especially in temperature variation, using features from deterministic crop models to get perfect statistical CO2 fertilization and consideration of fertilizer. The limited elaboration given on the three approaches for improving crop yield prediction is a weak side of the paper since better guidance could be given for any researcher that is doing work around this area especially due to the approaches being not explored to the level until this review was done.

Research Paper that was proposed by Surana, Rajswee and Khandelwal, Ritu [39] uses machine learning (ML) methods like Random Forest (RF), Adaboost, Gradient Boost, and Support Vector Machine (SVM) for yield prediction based on a dataset of 2201 instances found in it. To demonstrate the predictive ability of the data mining method used an extensive dataset, 5-fold cross-validation, and a Random Sampling model were used in this comparative work. The authors used a set of different measuring metrics such as false positive rate, true positive rate, recall, precision, ROC area, and F-1 score, for comparison of the performance that is achieved by the different classification algorithms that are mentioned above. From the results of these tests Random Forest algorithm was chosen as the best for having the best accuracy in comparison to the others based on the dataset that is used for the work. Similar to paper[32] this paper shares a weak point in its work in regards to the conclusion being based on non real-world agricultural contexts which makes it hard to accept the conclusion that is reached for the mere fact that it could change when it is actually applied.

2.13 Summary

Classifications, theoretical underpinnings related to time series and it's relation to yield prediction,research works done on areas comprising this , Prediction mechanisms, feature extraction strategies, remote sensing, weather data and attention mechanism were covered in this chapter.

Chapter 3

Methodology

The methods and strategies used to accomplish the research objectives are the primary contents of this chapter. The data pre processing is crucial for the use of machine learning models that are covered in depth in the first section. The actual procedure for creating and refining these models, including feature extraction methods and model designs is discussed in the next section. Finally, the chapter describes the method utilized to evaluate the models included in this work.Figure3.1shows our proposed model for the work.

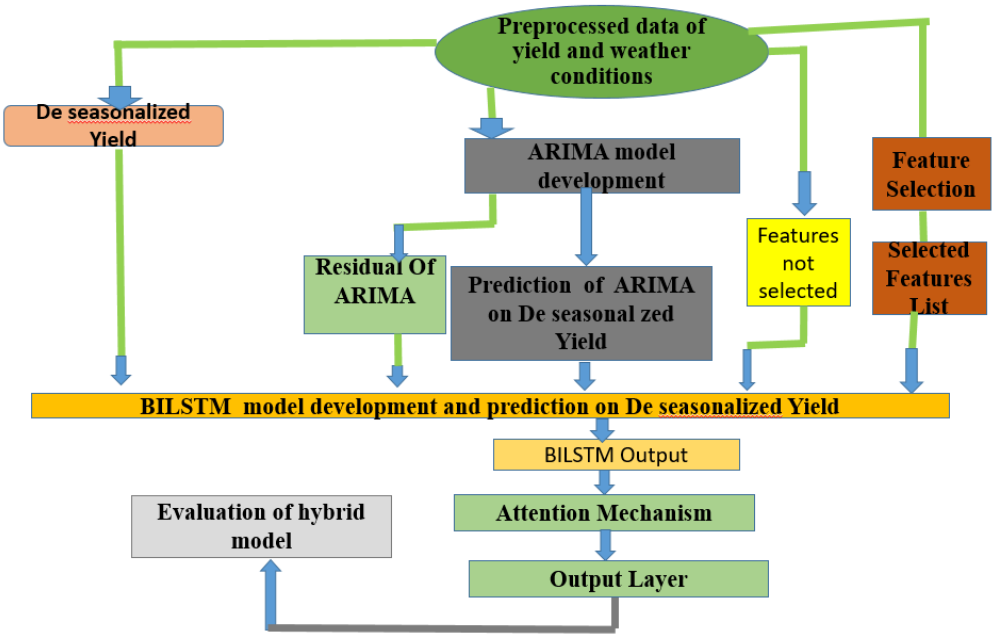

Figure 3.1: Proposed flow diagram of HYBRID ARIMA BILSTM model

3.1 Dataset for this work

Data that can be used for forecasting and yield prediction which is the objective of this work can be found in different sources. Despite the fact that it is a popular research area a dataset that includes remote sensing data along side weather and yield data is very rare and may not be freely available for use. The data set selected for this work constitutes these. We used the following criteria to select the dataset used for our work.

- If the dataset has wheat yield data for all the years the experiment took place for
- If the dataset contains weather data that were taken at different growth stages of the crop
- If the dataset contains NDVI values that is also taken at similar growth stages as the weather measurements
- If the dataset was free and available to use for any researcher

3.1.1 Dataset used for work

The dataset that is selected for this work is taken from a set of experiments that are used for a research paper done on the title of ”Simulation of winter wheat response to variable sowing dates and densities in a high-yielding environment” It included a collection of experiments that were taken at two sites namely Leeston and Wakanui in the Canterbury Region of the South Island of New Zealand.The first years from 2013 to 2015 experiment took place in the Leeston region while second part of the experiment took place at the second site from the year 2015 to 2018.The plot area for these experiments was taken to be 12*165 m for each plot.The weather station found two kilometers from the site of experiment contributed in registering and giving Vapour pressure, Max and Min temperature,rain snowfall and solar radiation at 2 m. While division of Trimble Agriculture Division, CO, USA contributed to the measurement values for the NDVI data. Accordingly the experiments included data for NDVI, Vapor pressure, Max and Min temperature,rain snowfall, date of measurement, growth stage ,solar radiation and plot number in addition to yield data for the years from 2013 to 2018.The dataset contains 3473 rows with 10 columns with the features which are collected from the months from February,early March, late March and April. Based on our criteria the following was out of many that checked all the criteria that was set it had the following charters tics that shown on Figure3.2

experiment ID	date of measurement	growth stage Zadoks	NDVI	solar rad	maximum temp	minimum temperature	rain snow fall	vapor pressure	YIELD
FARlee2015	04/09/2014	11	0.276017	3.561678	14.3	10.6	0.508	1.36	25450
FARlee2015	05/01/2014	21	0.270505	29.36463	23.58	10.48	10.48	1.22	25450
FARlee2015	05/09/2014	22	0.343227	4.760717	12.1	3.4	1.27	0.96	25450
FARlee2015	7/15/2014	29	0.816083	6.414564	11.8	-2.7	0.254	0.63	25450
FARlee2015	08/04/2014	30	0.815109	9.521434	10.1	-1.8	0	0.72	25450
FARlee2015	09/01/2014	31	0.791224	6.479537	11.5	4.6	0.254	0.94	25450
FARlee2015	9/26/2014	32	0.738966	9.574593	17.5	9.5	0.762	0.95	25450
FARlee2015	10/13/2014	33	0.713074	21.61814	24.8	8.2	0	1.14	25450
FARlee2015	11/04/2014	43	0.759223	16.63888	12.8	4	2.286	0.81	25450
FARlee2015	11/26/2014	70.2	0.743016	32.45675	25	9.5	0	1.01	25450
FARlee2015	11/28/2014	70	0.784863	16.30221	12.7	6.5	1.778	1.11	25450
FARlee2015	12/11/2014	79	0.731512	9.946709	13.8	8.5	0.254	1.12	25450
FARlee2015	12/23/2014	82	0.668285	27.08765	21.4	14.7	0	1.58	25450
FARlee2015	01/12/2015	87	0.192574	16.99427	24.55	7.764	0	1.65	25450
FARlee2015	1/20/2015	92	0.147675	27.64287	18.9	7.3	0	1.19	25450

Figure 3.2: Sample of Dataset

3.2 Pre processing

3.2.1 Dataset cleaning and filtering

The next step in the research work was pre processing the dataset so as to have a suitable and meaningful data for the machine learning model.Cleaning and filtering datasets are crucial steps in yield prediction since they greatly contribute to the accuracy, representativeness, and quality of the training data. Addressing problems with data quality, eliminating noise and inconsistencies, managing class imbalances, reducing biases, and standardizing formats are the main goals. The cleaning process aids in the production of a trustworthy dataset for machine learning model.

The following criterion's are used to filter and clean the above mentioned dataset that is used for the work.

- If the dataset has features included in the dataset that were simply there for labeling the data and have close to non correlation or relation that affects the target feature which is the yield in our case.
- If the dataset contains outliers that can not be handled for further usage during the modeling part of our work.

Upon these criterion's the data got refined and is used for the modeling work. The following are the different techniques that we used during this process:

1. **Data loading:** Initially we loaded the data to know the data type of each feature so that later on based on the data type conversion of data type is selected depending weather it is string,object, boolean or an integer.

2. **One Hot Encoding:** Based on the type of data we received on the first step we used is the One Hot Encoding to convert features having non numerical data type to numerical valued data types.

3. **Handling Missing Data:** After the conversion technique we handled the missing values found in certain columns through interpolation (e.g., using linear interpolation) to fill gaps in the data, which helps in maintaining continuity instead of simply dropping them.

4. **Date conversion:** The date of measurement column is converted to a date time format, ensuring that it is treated correctly in time series analysis.

5. **Removing NaN Values:** Any remaining rows with missing values that are beyond interpolation technique are dropped using dropna()

6. **Feature Scaling:** A Standard Scaler is applied to several numeric columns (such as growth stage, solar radiation, etc.) to scale them to have a mean of 0 and a standard deviation of 1. This ensures that all features are on the same scale, which is crucial for many machine learning models, especially for neural networks like LSTM.

7. **Seasonal Decomposition:** The seasonal decompose() function breaks down the time series into its trend, seasonal, and residual components. This helps in understanding the underlying patterns and removing seasonality for improved model performance. In order to do this part the time series is initially observed so as to know if the seasonal fluctuations occurring constant overtime or changing proportionally with the series.Accordingly the seasonal fluctuations appear constant in magnitude over time so we will decompose the series using an additive model using the following formula.

$$\text{ADDITIVE}(Y_t) = T_t + S_t + R_t$$

Y_t =observed value at time t

T_t =Trend component at time t

S_t =Seasonal component at time t

R_t =Residual component at time t

8. **De seasonalization:** After decomposition, the series is de seasonalized by subtracting the seasonal component from the original series, which helps in isolating the trend and residuals for modeling.

9. **Re seasonalization:** After making the predictions using the models the de seasonalized data needs to be transformed back to its original format so as to do a correct comparison between the original and the predicted one.

3.2.2 Data partitioning

Chronological data splitting is the correct choice when coming to data partitioning of time series data, This splitting ensures that the behavior of ordering is kept when doing the prediction work and would not be preserved if using random data partitioning. As result we use 70/15/15 chronological data splitting where the 70 is for training, where as the first 15 is for validation and the last 15 is for testing. And based on our dataset the 70 percent is form the month in range from April 2013- July 2016, the 15 percent for the validation is form the month in range from July 2016-March 2017 while the remaining 15 percent of the testing part is form the month in range from April 2017 -January 2018.

3.3 proposed model

3.3.1 ARIMA modeling

After the pre processing work of the data is done which includes the de seasonlization of the target variable through the step of decomposition of time series data ,the next step to making the first part of the HyBRID model which is the ARIMA model. Based on the fact that we have fluctuations that do not change with the overall level of the series, we choose to use ARIMA model for focusing on capturing trend and residuals. The seasonal component is regarded as part of the residual noise. After wards the chronological data split-ted three sections of the data the seasonal component is re indexed to match indices of these chronological splits.

3.3.2 Stationary Testing and Transformation

The next section of work in ARIMA modeling involves the following two steps

- Augmented Dickey-Fuller Test: This test helps in calculating the test statistics which is negative number and its value being more negative provides proof against the null hypothesis which would mean the series is non stationary since it has a unit root. The P value found in this test is a probability of observing if the test static if being extreme or more extreme when compared to the one that is calculated. Prior to conducting this test choosing a significance level which is the maximum probability of making the error of incorrectly rejecting a non stationary result or behavior. A common choice for this level is 5 percent (0.05) so in general if the p value is lower than 0.05 then the series is stationary and vice versa if it is not [40]. This test is checked and differencing is applied to the series until the test resulted in a series that is stationary with a P value lower than 0.05. The check for stationary is incorporated in the analysis for the acf and pacf
- Differencing: Differencing is applied to make the series stationary, as this is often required for ARIMA models and other time series forecasting methods. In order to meet the p value lower than 0.05 differencing is applied twice.

3.3.3 Auto correlation and Partial autocorrelation plot

Autocorrelation (ACF) is the correlation that shows how much a series is linearly dependent on its past values. While the Partial auto correlation (PACF)is the correlation between a series and its lagged version at a certain lag after the removal of the liner effects of the in between found lags.

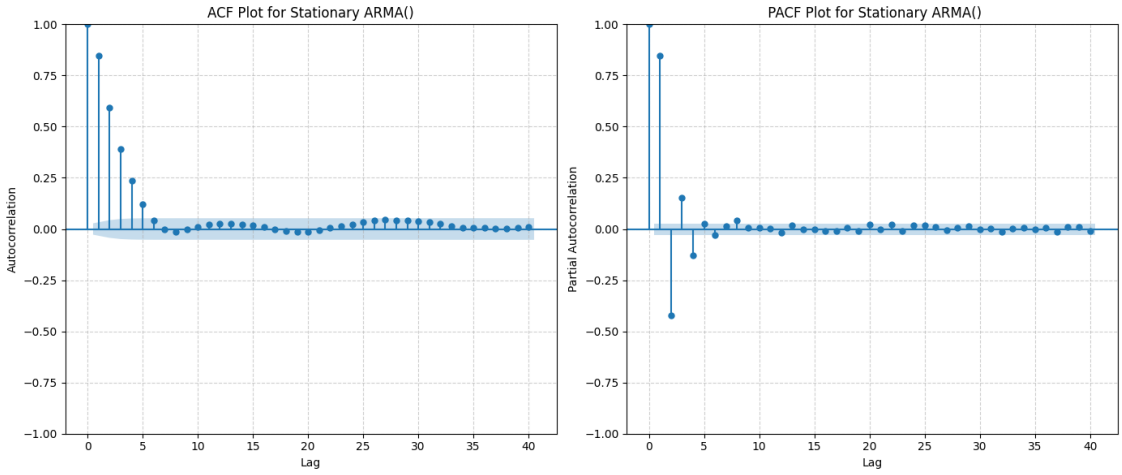

Figure 3.3: ACF and PACF plot

Based on the auto correlation and partial autocorrelation plot shown on Figure 3.3 the ACF shows a declining plot when going to the higher lags and it can be seen that after lag 6 the plot drops to zero. In similar manner when coming to the PACF plot found on the right where after the fourth lag the plot decays to zero.Accordingly the Auto Regressive Integrated Moving Average (ARIMA) parameters of order p,d,q are extracted.The auto regressive part of the model designated by the P from the above three letters will have an order that is based on partial autocorrelation function. The Pth value or order is set for a point that has significant lags which is not crossing zero. From this it was found that **Pth value in this work was found to be 4** The other parameter is the autocorrelation function which sets the order for the moving average part of the ARIMA model. It is designated by q letter and it will have an order where the lag becomes insignificant and **for this work the order is found to be 6**. The remaining part of the model is the differencing part denoted by the letter d and it depends on the series being stationary or not based on the earlier test of ADF we performed and as a result **for this work it was found that d had a value of 2**. After the order for the ARIMA model is set the next step is to find prediction done by the model on the de seasonalized series for training ,validation and testing part of the data based on the chronological data split.Which is later on followed by adding back the seasonal component for making the predictions in the original scale and then evaluate the predictions against the actual using the metric evaluations.

3.3.4 The BiLSTM Model

The BiLSTM model is used to compensate for the shortcomings of the ARIMA model that cannot be dealt well when doing the yield prediction but prior to developing the next part of the HYBRID model which is the BILSTM features that are going to be used as an input for it need to be initially selected , the attention mechanism needs to be selected, the optimizer selection process also need to be established and they are done through the following process:-

3.3.4.1 Feature Selection process

1. **Descriptive analysis:** For the selection of features one analysis we performed was the descriptive analysis where all numerical feature's distribution is seen across the different periods of the data along side the features mean,median and standard deviation presented in histograms. In addition, this analysis also helped us to see how many NAN values that are still existing in our data even after the handling of them.

2. **Correlation analysis vs YIELD:** Since all features can not have similar effect on the target variable which is the yield next we proceeded to do correlation analysis of every feature with the yield and based on this analysis the selected features were selected and kept for a further selection process.

3. **Granger Causality:** We used this analysis test to understand if past values of one feature can help predict future values of another feature. In our case it is the features and how much they affect the yield.This is applied through the following code shown on Figure 3.4

```
# 2. Granger Causality
maxlag = 1
for feature in all_potential_features:
    if feature in df.columns:
        try:
            data = df[[target_column, feature]].dropna()
            if len(data) > maxlag + 5:
                gc_res = grangercausalitytests(data, maxlag=maxlag, verbose=False)
                p_value = gc_res[maxlag][0]['ssr_ftest'][1]
                feature_importance_df.loc[feature, 'Granger_P_Value'] = p_value
                if p_value < granger_p_threshold:
                    feature_importance_df.loc[feature, 'Combined_Score'] += 1
                    if feature_importance_df.loc[feature, 'Selection_Reason']:
                        feature_importance_df.loc[feature, 'Selection_Reason'] += ', '
                    feature_importance_df.loc[feature, 'Selection_Reason'] += f"Granger_p<{granger_p_threshold:.2f}"
            else:
                print(f"Warning: Not enough data for Granger test on {feature}")
        except Exception as e:
            print(f"Error in Granger test for {feature}: {e}")
```

Figure 3.4: Granger Causality

Initially the test will check if one past value of the feature helps in predicting the yield value this is then followed by iteration through the different features found in our dataset. The length for each individual feature is checked so that test is skipped if the data length is small or short.During the application of the test the function grangercasulaitytests based from the statsmodels library in python is called and in return it results a dictionary where keys are the lag numbers. Each lag's result is a tuple where the first element is is the F test result while the second element is the P value. This P value is stored in (feature importance df) under the Granger p value column for the current feature.We have used granger threshold of 0.05 considering the null hypothesis theory of past values actually influencing current target values. So based on this if p value is less than threshold then it means that the feature does Granger-cause the target variable and this also means that past values of that feature help to predict the yield's future values.

Since these tests mentioned so far are applicable for numerical valued features we used other analysis and feature selection methods that also deal with the non numerical features that are found in the dataset. The following are some of them

4. **Mutual Information:** This test helped us to understand one feature's characteristic through another feature regardless of what their data types were. And its applicability for both numerical and non numerical gives our feature selection an augmentation to its work.

5. **Random Forest Feature Importance:** Another analysis that is used also for both numerical and non numerical features is the Random Forest feature importance, The application of this model shows us the predictive influential power each feature has on the target feature which is the yield in this case regardless of them being numerical or not.

6. **Combined Score:** In this section a function is built that tracks the results that is obtained by the above five tests and then a combined score is created which increments to one whenever a feature passes any of the above tests set thresholds. Finally the function considers all the features that have obtained a combined score above 0 and as a result it is the selected final features are displayed to the console window.

3.3.4.2 Attention Mechanism Selection

The BILSTM model has the de seasonalized target variable, the selected features,ARIMA prediction and residual that are done on the de seasonalized target and original data.The model processes these inputs and results in an output that is fed to the attention mechanism.An attention mechanism is applied so as to help the BILSTM model in making a correct decision by giving it areas in which it needs to focus on.Its selection process is done through using suggest attention mechanism function which checks several things and later on gives the suitable attention type for the work at hand. The different checkups done on the time series are the following.

- **Input for the Function:** This function checks acf pacf results,feature importance based on correlation,the selected features chosen from the feature selection section,cleaned data frame along with columns of date and yield.
- **Suggest Logic:** The function suggests a suitable attention type by doing conditional checkups in the following way.
    1. **Strong Autocorrelation:** If differencing (d) found from the result of acf pacf shows a value greater than 0 and indicating a strong autocorrelation which implies a temporal dependencies which is the yield variable and thus applies self attention multihead attention mechanism. Otherwise it will suggest other attention type that are much more simplier like the self attention type.
    2. **Existence of many features:** If the selected features that resulted are alot in number in our case we have taken 7 as our base number and greater than 7 then the attention type suggested would be in this section an attention type that is capable of dealing with complex relationships not just with in the time steps of a single feature but also between multiple features. But if number of the selected feature is less than 7 then an attention that deals with lesser complex relationships will be suggested.
    3. **High correlation with most important feature check:** If a single variable or feature is highly dominate in affecting the yield feature then it suggests to use a single head self attention mechanism otherwise the checkup accepts the attention type that is selected by the other checkups.

4. **Frequency Domain Analysis:** In this step the function checks for seasonal and cyclic patters that are existing in the yield data and tries to find a dominant or strong frequencies in the series which is done through applying analyze frequency domain. If mid range frequencies are found the recommended attention is a simple attention mechanism since it would benefit it to focus on its self. And if high frequency is found meaning rapid and short term patterns it would suggest the no application of an attention mechanism since a simpler model without an attention will help it than one that has. But if both scenarios are not detected then and nothing is suggested and would accept the suggestion for the attention mechanism by the other attention mechanism selection analysis tests that are mentioned earlier.

The types of attention mechanisms we have included for this work as choices are the simple (additive) and Multi head attention mechanism. In our case the selection function going through the above selection process resulted in choosing Multi head attention mechanism as the appropriate one for the model. How this mechanism helps the BILSTM in achieving a good prediction is described in the following way. In order to decrease the complexity of the model we choose the BILSTM output having the shape of (batch size,look back,2*LSTM units) to be passed as an input for the Query(Q),Key(k)and Value(V) making it a multi head self attention.Initially the BILSTM output is passed to the number of heads that we have chosen (in our case 2 is the number of heads)to create distinct Query,Key and Value matrices where these transform 2*LSTM units into Key-dim which is equal to the lstm units in our case, this accordingly will result in both heads having the shape of (batch size,look back,lstm units). Next raw scores are calculated as dot products of Query and Key for each pair of time steps. After this the softmax function gets applied in order to normalize these scores into proper weights that are later on used by for the making of the weighted sum.This weighted sum in return will help in making the context matrix for each head and after finding the individual context matrices they are concatenated and linearly transformed again to produce the final output of the multi head attention layer having the shape of (batch size,look back, 2*lstm units). This is later on fed to the flatten layer which reshapes it to a 2D creating a single long vector for each item in the batch. This reduced shaped context vector is then passed directly to the dense layer where the layer performs a final liner transformation on the combined features to produce the model's prediction for the target variable.

3.3.4.3 Optimizer Selection

Another component to work is the optimizer used and in its selection process we have taken into an account that for an optimizer to be termed good, it ought to have a behavior of converging early,obtains a low training loss, generalizes well to unseen data, has a stable training process and is not overly sensitive to hyperparameter. So with all these kept in mind we choose to iterate through the three most commonly used optimizers which were Adam, RMSprop and SGD. The way how the iteration goes is done among these three optimizers is due to the evaluation of optimizer function we had setup. This function creates an instance for each of the optimizers depending on the if statement asking for lower cased name of optimizer. So when the evaluate model with optimizer function is called the weights are updated by all three optimizers and the validation loss achieved by each of them is registered and optimizer that archived lower is selected for retraining of the model. The following are parameters used in the structure of the BILSTM

- Units: 100 units in each of the BILSTM layer totaling to 200, which control the number of output features and the capacity of the model to learn patterns in the data.
- return sequences=True: This ensures that the BiLSTM outputs a sequence value, which is used for the attention mechanism as an input.
- Dropout Layer: A Dropout layer follows the BILSTM to regularize the model. Dropout randomly sets a fraction of input units to zero during training (in this case, 30 percent of units are dropped out). This helps prevent over fitting and improves the model's generalization ability.
- Plotting BiLSTM Results: Similar to ARIMA, the predictions of the BiLSTM model are plotted to assess its performance over the training, validation, and test sets. The obtained prediction are therefore used as also the output of the hybrid model.

3.3.5 Hybrid model Implementation

HYBRID ARIMA-LSTM : The input data initially is in a 1D array format for the ARIMA model (n samples) which is also the same dimension of its output(n predictions).The prediction and residual outputs of the ARIMA are generated on the next stage. In order to do proper comparison with the actual yield value the ARIMA prediction is re seasonalized and plotted. Next we prepared dataset for the BILSTM containing the de seasonalized yield ,selected features , the ARIMA prediction and residual. This data was structured into a sequence through using a look back(previous time steps to consider for predicting the next) of 18. This step leads to having a 3D input of a shape (batch size,look back, n features) , where n features is yield + number of selected features + prediction and residual of ARIMA. The output of the BILSTM are hidden states for each time step in the input sequence having a shape of (batch size,look back, 2* LSTM units). Where the appropriate LSTM unit we found for work is at 100 and 32 for the batch size which makes the output shape as (32,18,200). After this step this output is fed to the multi head attention mechanism (the suggested mechanism for the work) which has an input shape same as the output of the BILSTM of 3D. The attention mechanism gives weighted sum for all the look back time steps and after the results of all the heads is concatenated (where number of heads is 2 in our case) its output shape becomes (batch size,attention output dim) where attention output dim is 2*LSTM units. This output is a context vector which is fed to a dense layer which in return has an output shape (batch size,1)

3.4 Baseline Models

The proposed model should be verified by other models that can be used for the work of yield prediction. we chose the baseline models of BILSTM without attention mechanism and ARIMA.

3.4.1 BILSTM without an Attention Mechanism Model

This baseline model also is structured as having no input from the ARIMA model and also not having the attention mechanism. This baseline model also helps in to show the effectiveness of the attention mechanism that is applied to the HYBRID model.In similar manner the BILSTM for this baseline model was also initially programmed the same as the HYBRID one but through different experiment no of LSTM units being 200,patience of 20 and L1,L2 both being 0.001 gave a better resulting model.In addition the optimizer selection function is also applied for this model. The optimizer selection goes in parallel with the complexity of a model.The flow model is shown on Figure 3.5

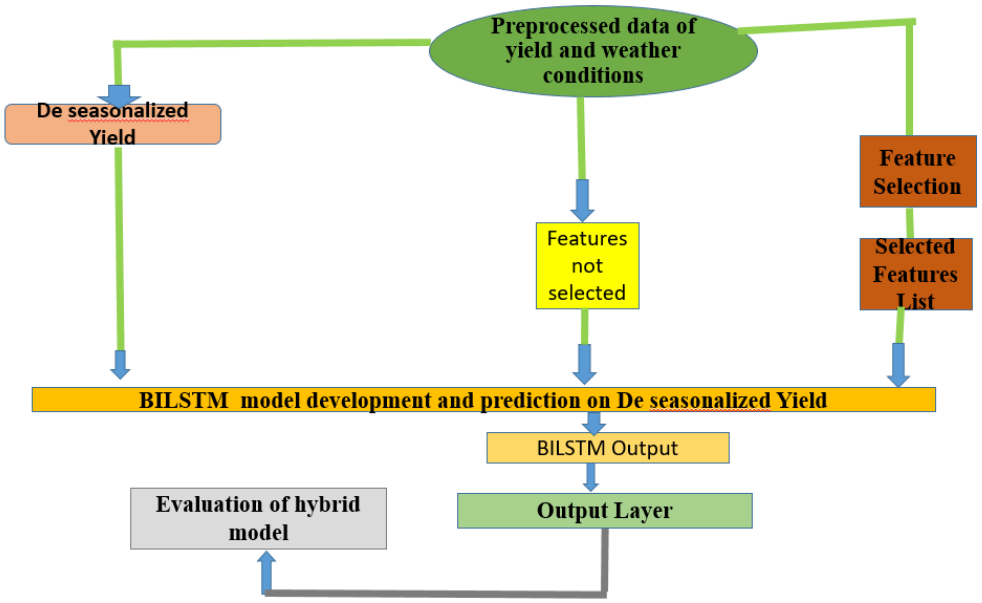

Figure 3.5: BiLSTM Without Attention Mechanism model flow diagram

3.4.2 ARIMA MODEL

ARIMA is one of the oldest prediction models and for our work we are using 4,2,6 for the values of p.d.q.Using the ARIMA model as a baseline model helps us in understanding how much the hybrid model's performance has improved in comparison to a simpler model like the ARIMA.

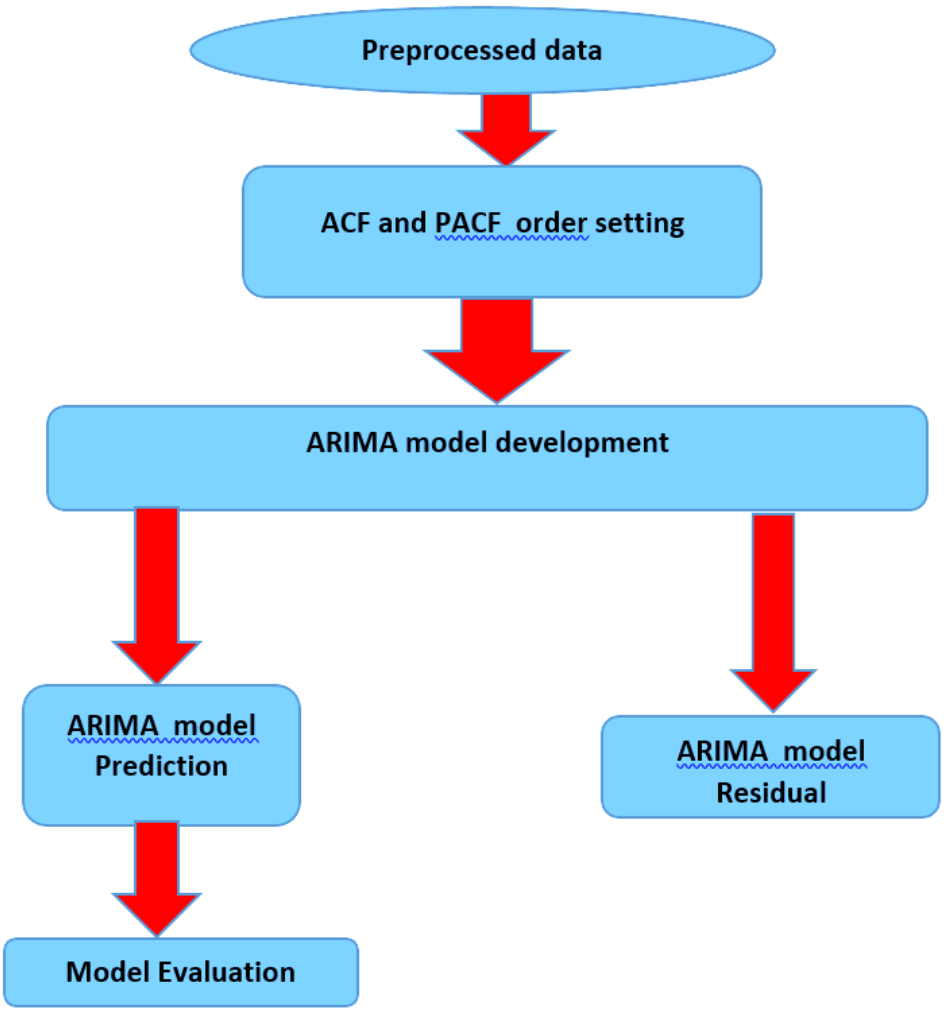

Figure 3.6: ARIMA model flow diagram

3.5 Alternative Model Approaches

3.5.1 BiLSTM With attention Model

The BiLSTM With attention is used as one of the alternative models for performance comparison and is structured as having no ARIMA prediction and residual but uses the attention mechanism that we had applied to the HYBRID model. From this comparison we aimed to see how much the attention incorporated HYBRID model performed and if its effectiveness is also repeated on the alternative model.The same selection process for the attention mechanism and optimizer selection is also applied to this model.For the BILSTM architecture initially we used similar parameters as it was done for the HYBRID model but through various experimentation It was found that no of units being 32, patience of 20 and L1, L2 being 0.001 gave us a better performing model.Figure 3.7 shows the models flow diagram that we used for our work.

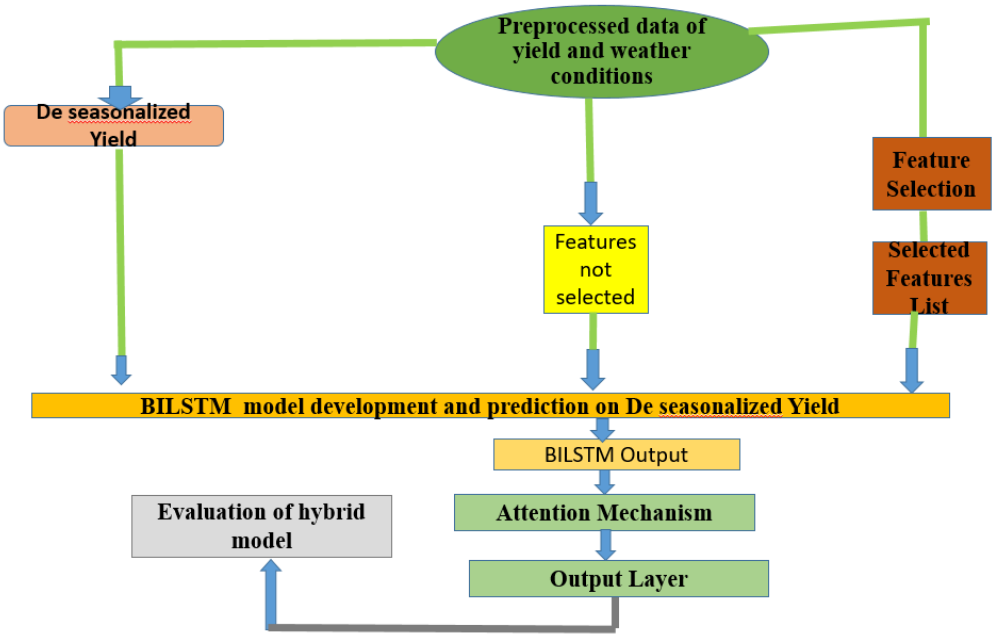

Figure 3.7: BiLSTM with Attention model flow diagram

3.5.2 Window Sliding MODEL

The Window Sliding MODEL helps in understanding the relationship of both ARIMA and LSTM models when they are applied to a certain dataset. How both models cope with the actual value found in the dataset without the addition of any other feature to help in the prediction work proves the theoretical idea that is said about the two models which is in the case of ARIMA the inability to deal with complex and non liner patterns and having capacity of long term memory that helps in dealing with non linear parts of data in the case of BILSTM. So in order to understand this relationship another model we applied for this work is the window sliding model in which ARIMA is fitted to the data and outputs a set of results up to a point where the residual gets large enough that ARIMA no longer can properly predict and that point is called the failure point. After the ARIMA prediction deviates from the true values,the BILSTM starts its prediction work from the failure point onward compensating for ARIMA's inability to deal with complex and non-linear data points.The flow diagram shown on Figure 3.8 shows the general steps involved in this work.

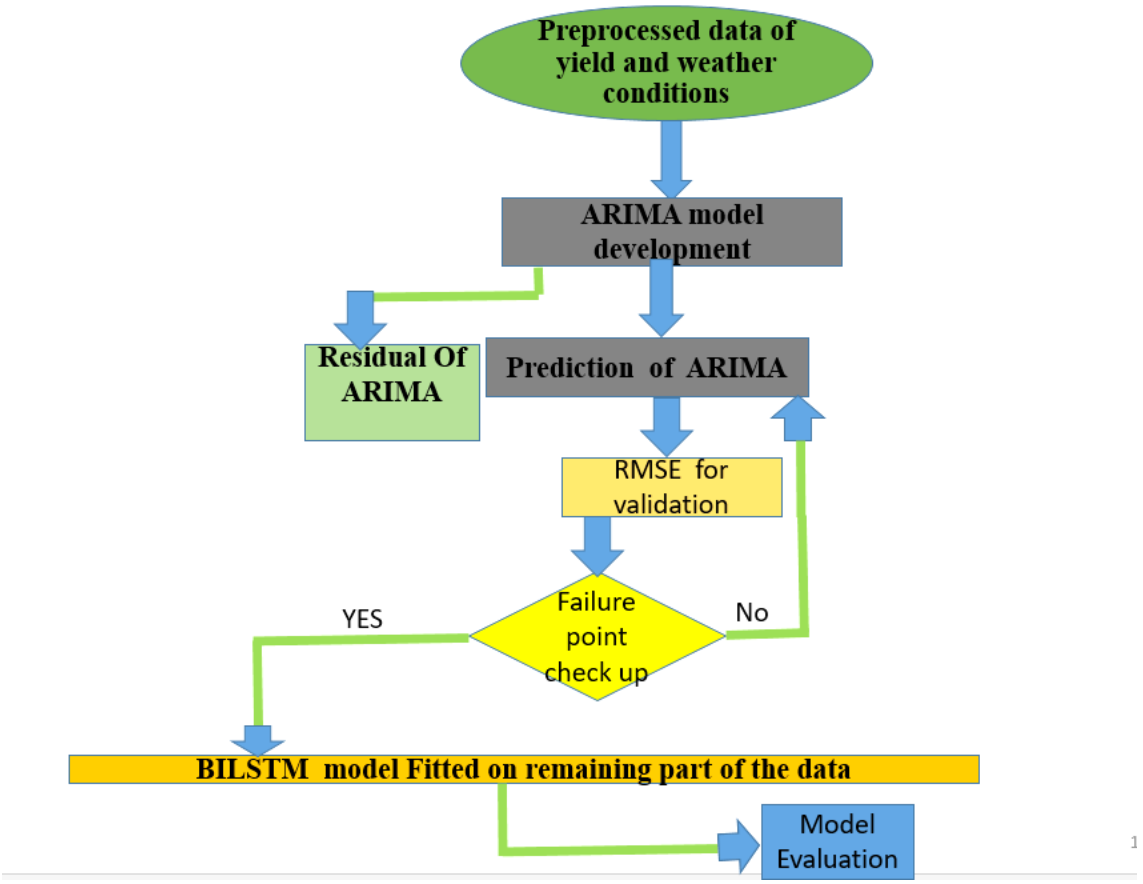

Figure 3.8: Window Sliding MODEL flow diagram

3.5.3 Window Sliding MODEL implementation

This model is applied by combining the following three parameters which are the failure threshold setting,the window sizing and the step size. The following steps are what we followed to do the experiment for this model.

1. Chronological data Splitting:- Data is initially split in a chronological order so that the model also learns temporal behavior of the data.

2. Failure threshold check up and setting:- How a failure point is set for a certain work has the ability to affect the outcome of the work since if set to a high valued number it ignores deviations that occurred before it, leading to the BILSTM model will start prediction work very late and vice versa when it is set to a low value. So instead of using a fixed rate for it using a dynamic changing value helps the BILSTM on all the failures that occurred due to lack of prediction from ARIMA. In addition to this the threshold setting mechanism also has to be robust enough for outliers in being able to capture real tendency of generated errors. This process involves the following steps

   (a) ARIMA prediction and RMSE value:- ARIMA model is fitted to the training dataset and both prediction and residual are outputted for both the training and validation data sets.Afterwards RMSE is calculated for the training and validation predictions that were generated previously.

   (b) Failure thresholding:- This value is set next by doubling the validation RMSE that we achieved earlier.

   (c) Rolling window setting:- After absolute error between the actual and predicted are calculated we then applied the rolling window to calculate mean of the absolute errors. In addition we specified 3 consecutive failures being above the threshold to be used for failure detection.

   (d) Failure point setting:- If a failure point is detected based on the the rolling error exceeding the threshold then it is logged and shown on the plot as a ARIMA failure point otherwise the code returns a message saying no failure point detected.

This being said for this work we made use of **metric results of RMSE** which are the results of validation part of ARIMA as a reference. Result of the Any deterioration beyond double the value of RMSE for the validation value will be defined as a failure point. We also added a rolling window to our model so as to make failure detection more robust and not to consider all spikes in the data as a error.

3. Window sizing:- It is the practice of analyzing a subset of the entire dataset through a certain window within a specific observation. It helps in defining the amount of data that could be included within the subset chosen for analysis for making a prediction.Selection of this size should consider amount of data.We experimented our rolling window with the values of 5,7,10,14 and finally achieving a good result with a rolling window value of 5. We used this rolling window to do mean of the absolute errors over the last rolling window data points.After this the average of last rolling window data points is checked if it is above or below the already set threshold. Based on its result if checked to be higher than set baseline then BILSTM starts its work.

4. Step Size:- Another parameter that we made use of is the step size which helps in determining how much the window is moved for the next iteration A step size 1 is taken for our work for better understanding of the work after trying step sizes 2,5 which resulted in a failing to get outputs before the window is shifted to the next iteration.

3.6 Performance Evaluation Metrics

Evaluation metrics for yield prediction should be based on the nature of the data,the metrics ability to be easily interpretable,its capability to help models focus on the important deviations that are seen instead of all, and its capacity to be adaptable to the nature of the data. and based on these criterion MSE, MAE, RMSE, and MAPE are what we used as performance metrics to assess how much the predicted value deviates from the true value.We used these metrics to evaluate the hybrid model approach that we are proposing.

1. **Mean Absolute Error (MAE)** for computing the average absolute differences between predicted values and actual values

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

y_i =actual value

$\hat{y}_i$ =predicted value

n =number of data points

2. **Mean Squared Error (MSE)** for computing the average of squared differences between predicted values and actual values

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y_i =actual value

$\hat{y}_i$ =predicted value

n =number of data points

3. **Root Mean Squared Error (RMSE)** for measuring the square root of the average of squared differences between predicted values and actual values

$$\text{RMSE} = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

y_i =actual value

$\hat{y}_i$ =predicted value

n =number of data points

4. **Mean Absolute Percentage Error (MAPE)** for computing average percentage difference between predicted values and actual values

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

y_i =actual value

$\hat{y}_i$ =predicted value

n =number of data points

Chapter 4

Result and Discussion

In this chapter we will discuss about the experiment setup,the tools that are used to complete the experiments, the results found after the experiments and discussion for experiment result.

4.1 Experiment Setup

The experiments were conducted using Python 3.9.16, TensorFlow 2.17.0 , and Keras 2.12.0, installed on a HP envy laptop. This laptop is equipped with an 11th Gen Intel(R) IRIS Xe Core(TM) i7-1165G7 @ 2.80GHz 2.80 GHz, running the Windows 11 operating system.

The following are software and tools used in the research experiment:

- **Python**: is high level and general-purpose programming language that has abundance of libraries and frameworks that facilitate coding and save development time.
- **TensorFlow**:is a free and open-source software library for machine learning and Artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural network.
- **Keras**: is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.
- **NumPy**: is a robust numerical computation package for the Python programming language. Large, multi-dimensional arrays and matrices are supported, and a number of mathematical operations can be performed on these arrays. A key library for scientific computing in Python, NumPy is extensively utilized in domains like data analysis, scientific research, and machine learning.
- **Panadas**: is a prominent and powerful Python programming language and data manipulation and analysis toolkit. Along with tools for data cleansing, exploration, and analysis, it offers data structures for effectively storing and handling huge datasets.

- **Scikit-learn**: is a well-known machine learning library for the Python programming language. It offers a variety of machine learning tools and utilities for tasks like classification, regression, clustering, and dimensionality reduction, along with easy-to-use and effective tools for data analysis and modeling.

4.1.0.1 Pre processing implementation

1. **Dataset Loading** the first step in the implementation of the model was to load the Yield,weather having dataset to the python environment.We used the pandas library to load the data.

2. **Dataset cleaning and pre processing :** The loaded dataset was cleaned for better performance of the model by using the following methods

    - Irrelevant column dropping :- for removing columns that do not have any contribution for the prediction work.
    - Non numeric column conversion:- Non numerical values of features are converted to numerical valued features using one hot encoding.
    - Handling missing values:- Interpolation and removing technique are used for data that incomplete.
    - Organizing data chronologically :- Time dependent analysis accuracy depends on this part majorly as a result the data is sorted by the date.
    - Data spiting:- The training,validation and testing data split in a 75,15,15 percent proportion.
    - Scaling Data:- Standard scaler, Log transformation and inverse transform are the techniques that are used for scaling. The standard scaler helps in transforming the data to have mean of 0 and a standard deviation of 1 while the log technique helps in reducing skewness in the data.Inverse transform which helps in transforming predictions back to their original scale after standard scaler is applied.
    - Date conversion:- helps in converting dates not in the correct format
    - Stationary check up:- Application of the augmented Dickey -Fuller test for checking the difference series.
    - Seasonal Decomposition:- Makes use of perform timeseries decomposition function for doing this work while De seasonalization uses seasonal function for removing seasonal effects from the target variable

To proceed with building the hybrid model , the relationship among features and their relationship with the yield variable found in the dataset must be known. For such reasons the correlation matrix analysis is done on the next stage of works and discussed in the following section.

4.1.0.2 Correlation Matrix Result

As it can be seen in Figure 4.1 growth stage and normalized difference vegetative index have a highly negative relationship with the target variable and solar radiation and minimum temperature has weak positive correlations with the YIELD. Based on this findings we believe it would be logically sound to have a feature importance plot that assigns higher weights to these features.

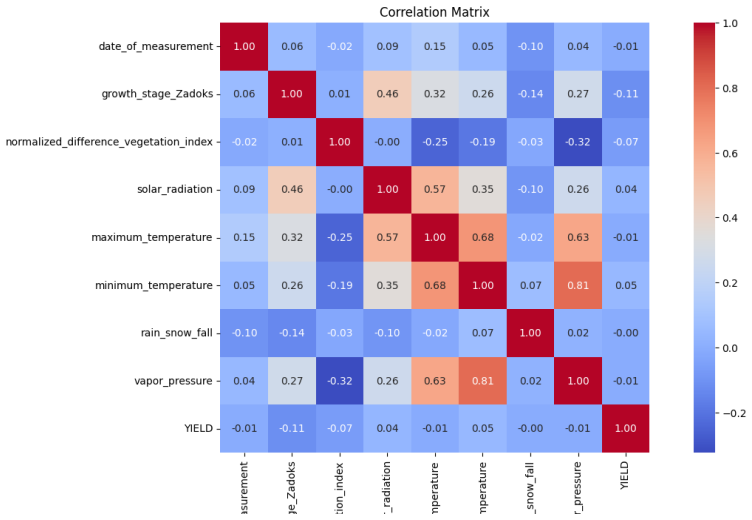

Figure 4.1: Correlation matrix

Figure 4.2 shows the feature importance weight plot for all the features found in the data set based on Figure 4.1 showing relationship of features with the yield variable.

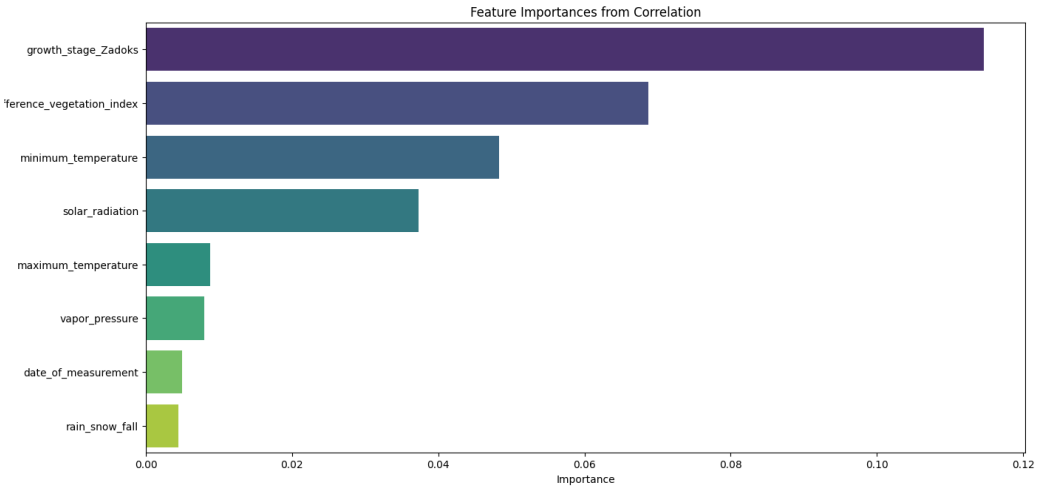

Figure 4.2: Feature importance

In order to see which numerical features that are found in the data that are highly affecting the target variable we performed the correlation matrix for only the numerical features which is shown in the heatmap shown on 4.1 and shown on greater depth on the feature importance on 4.2 which is also done for the numerical features. These both show growth stage,normalized difference vegetative index,minimum temperature and solar radiation are assigned to higher weights in comparison to the other features. Followed by maximum temperature, vapor pressure, date of measurement and rain snow fall. In order to see the feature importance for all numerical and non numerical features which can affect the yield highly is shown on 4.3. The features taken from these are passed to the other tests so that a final combined set of selected features is actually used for the modeling work.

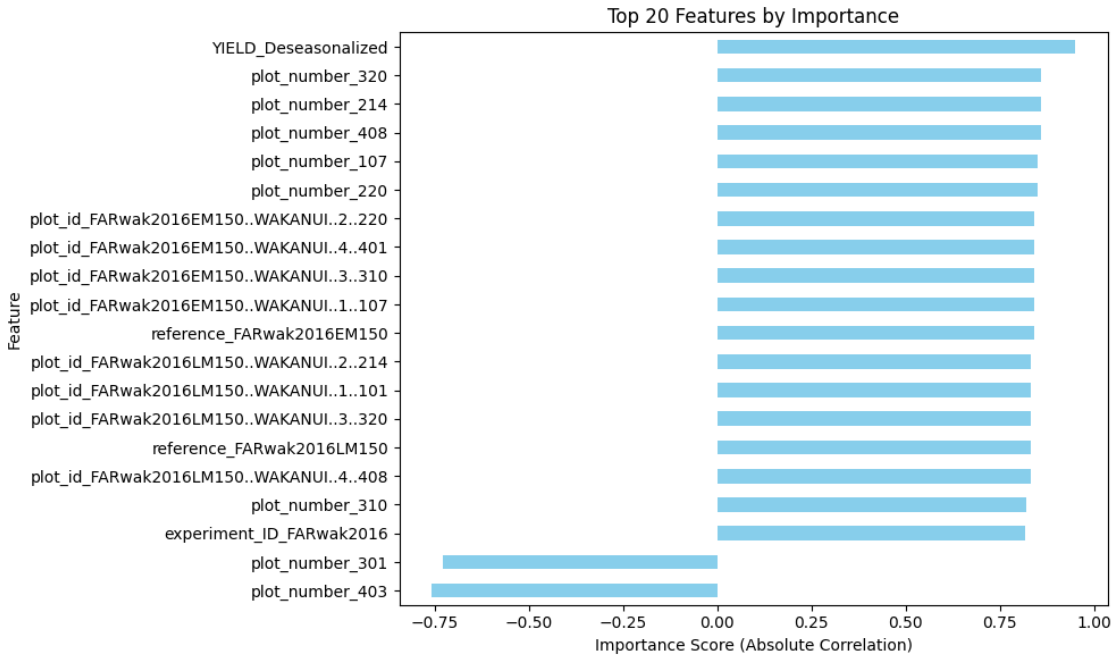

Figure 4.3: Feature importance for top the 20

Both the feature importance figures shown on Figure 4.2 and 4.3 are obtained by using the Pearson Correlation Coefficients having the following formula.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

- x_i and y_i are individual sample points
- $\bar{x}$ is mean of x values
- $\bar{y}$ is mean of y values
- r is Pearson Correlation Coefficient
- n is number of data points

The experiments done for the model involve using the following five hyper parameters and tuning them interchangeably on the dataset.

1. Batch size :which is the number of samples that will be propagated through the network.In our experiment we used 8,16,32,64 batch sizes interchangeably.
2. Epochs: refers one whole iteration of the training dataset through the learning process.Different number of epochs were used in our experiment such as 50,100,250

3. Drop rate: is a data, or noise, that is purposefully removed from a neural network in order to speed up processing and result turnaround And for the different experiments that take place different dropout rates are tested the values are 0.2,0.3 and 0.4

4. No of LSTM units : Different number of layers are also considered for this work and among the number of layers the experiments took place on the following 50,100,150,200 were used on both layers of the BILSTM model.

5. Patience :- Early stopping is also applied for this work and correct amount of early stopping was also checked among the different experiments. The choice of this parameter affects the model if not done since the model might go on learning noise in addition to the data.Some of the experimented values for it are :- 20,50,100

6. Regularization:- L1 and L2 are also applied to prevent overfitting and help the model generalize well to the actual data. Different values for these two are set among them are 0.1,0.001,0.0001 and the best value that was part of achieving the accepted result is when both were set to 0.1. The Multi head attention mechanism is the selected mechanism among the single headed additive and multiplicative attention types based on it fulfilling the requirements set by the selection function and the optimizer selected for this work is Adam which resulted in a better validation loss of 3.7475 in comparison to the Validation Loss achieved with sgd which is 33.7323 and 14.0341 with rmsprop. The following parameters used for the work are used on a fixed range and to mention some of them they are:-

The best result based on the different metric tests was obtained using Adam optimizer , batch size of 32, on 200 epochs, L1 and L2 being 0.1, No of LSTM units being 100 and at a patience value of 70

4.2 HYBRID model performance

In this section, the performance of the hybrid model is shown through prediction plots and metric test results how much the whole hybrid model understands the original data. Accordingly, the initial experiments done for the Hybrid model involve knowing the prediction result of the ARIMA model followed by the BILSTM models which is also the Hybrid models prediction result.

4.2.0.1 The ARIMA Model

- **ARIMA Model** : Stationarity of the date is a major step prior to fitting a model to the data so in our experiment we have applied differencing twice due to initial Augmented Dickey-Fuller Test showing non stationary behavior. After the second differencing the test shows a vaue of confirmation of stationary through the values of ADF statistic = -3.1306 and P-value of = 0.0244. And based on the critical value which is set as a bench mark for evaluating to either accept or reject the null hypothesis of a unit root in our data. The test set at 5 percent equating to a value of -2.860 where our experiment is greater than the set bench mark the test showed us the stationarity of the series so as to proceed to next experiments on our work which is fitting of the ARIMA model. The ARIMA model is trained on the deseasonalized data mentioned above with a specific hyperparameter (order = (4, 2, 6) and generates predictions on the training, validation, and test sets. The residuals (the differences between the predicted and actual values) are then calculated and also used as inputs later on.

- **Plotting ARIMA Results** :- Predictions from the ARIMA model are plotted to visualize how well the model fits the data and to check for any remaining patterns in the residuals. After the prediction and residual of the ARIMA generated they are concatenated with other input features so as to create an input to the BILSTM model which in return has an output that is used as the input for the attention mechanism.

ARIMA model prediction is shown on Figure 4.4 having all three data sections which are the training,validation and testing part of the data combined together for better visualization and understanding of the model's ability.

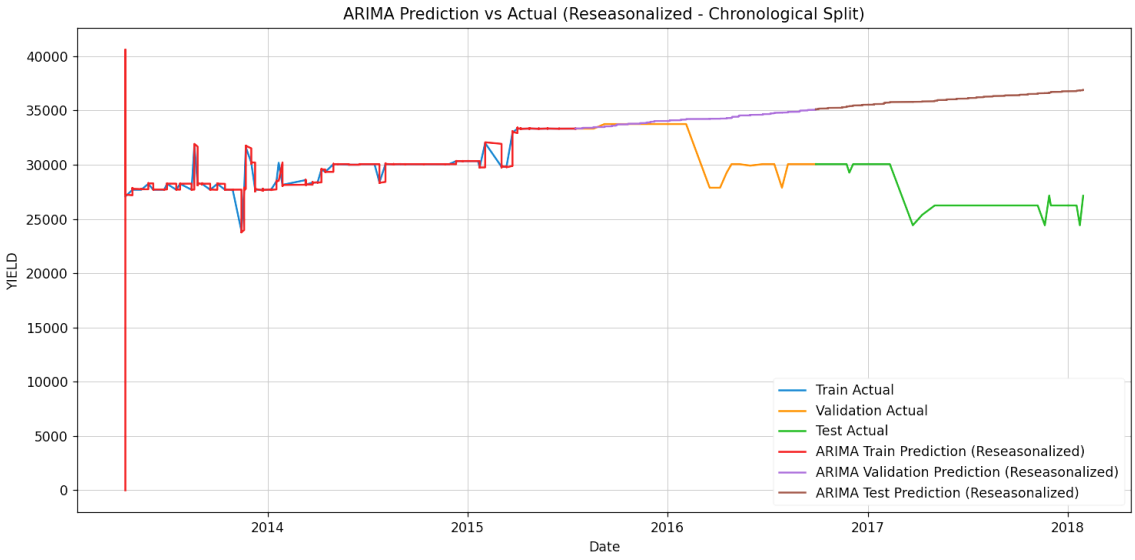

Figure 4.4: ARIMA Model prediction plot for Training,Validation and Testing

From Figure 4.4 it can be seen that starting from the validation part of the data, the ARIMA model fails to predict the data starting which starts from the end of 2015, clearly showing the need for a model that can predict for the rest of the data. Due to this reason the testing part of data that the model used shows only the original time series and a non-visible prediction plot where in addition to this the residual plot still shows a rising graph which shows that the model is failing in capturing the essence of the true data. The results of the metric tests for the ARIMA model are the following shown on Table 4.1.

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	646264.29	11774636.55	75861640.22
MAPE	0.28	8.38	31.11
RMSE	803.91	3431.42	8709.86
MAE	78.25	2507.82	8393.06

Table 4.1: ARIMA model metric results

4.2.0.2 The BiLSTM Model

The second part of the HYBRID model is the BILSTM part and Figure 4.4 shows that the majority of the original series is not dealt well by the ARIMA model and by applying this BILSTM model the remaining part of the original series will be put to use. The BilSTM model has an input of ARIMA prediction, ARIMA residual, list of selected features and original time series.The model processes this and results in an output which is fed to the attention mechanism which takes the output as a query and value and does attention scores followed by softmax operation for obtaining weights. It uses this weights for creating a context vector that is used for capturing most relevant parts of the BILSTM's output. This output is a context vector which is fed to a dense layer which in return has an output shape (batch size,1).

The best result based on the different metric tests was obtained using Adam optimizer , batch size of 32, on 200 epochs, L1 and L2 being 0.1, No of LSTM units being 100 and at a patience value of 70 the overall HYBRID model will have the following prediction plot. The green dotted line represents a demarcation line between training and validation data parts while the second dotted line shows the demarcation line between the validation and test data split ted parts.

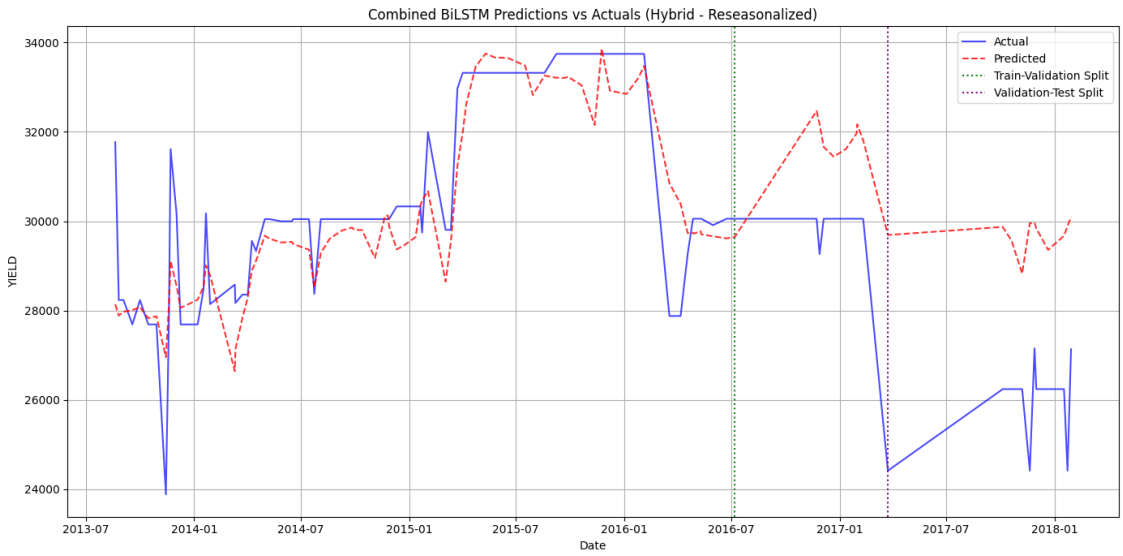

Figure 4.5: HYBRID model Prediction plot

As it can be seen from Figure 4.5 the HYBRID model has shown great improvement compared to the ARIMA model in understanding the original time series data. The prediction test plot clearly shows this conclusion. The metric results done for this model also show similar result as they are the following ones shown on Table 4.2.

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	129660.48	4922923.06	1356620.65
MAPE	0.79	5.76	3.13
RMSE	360.08	2218.77	1164.74
MAE	230.27	1740.85	887.23

Table 4.2: HYBRID model metric tests

From the metric results shown it can be seen that the HYBRID model is still overfitting despite many changes in parameter were done for it. Even though it shows progress on the test data the fact that during the validation its results being high raises a question in its ability to cope up. The experiments we did in regards to correct this situation have numerously gave us metric results that are worse of than the ones shown here.

4.3 Comparison of proposed model with baseline models

4.3.0.1 BILSTM without Attention mechanism

The first model we experimented performance comparison with the HYBRID is the BILSTM without attention mechanism. This model also shows us if simpler models are much better for prediction work when coming to our dataset. The resulted prediction plot of this model is shown in Figure 4.6 along with metric results obtained shown on Table 4.3. The Dotted line around mid of 2016 and earlier part of 2017 are demarcation between training and validation with the green dotted color line while the purple dotted line is for dividing validation from the test part of the data. The prediction is achieved at epoch 200,batch size 32, lstm units of 32 for each layer,L1 and L2 each with a value of 0.01 and SGD as an optimizer with a validation loss of 0.2868.

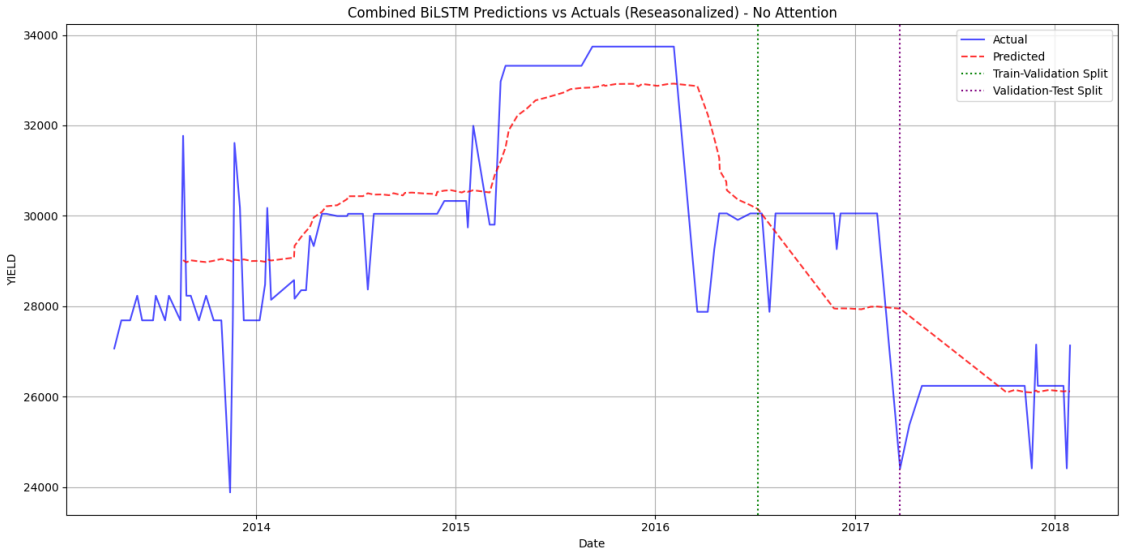

Figure 4.6: Prediction plot for BILSTM without Attention Mechanism

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	1913625.63)	5154084.15	786282.59
MAE	1040.38	2213.82	647.77
RMSE	1383.34	2270.26	886.73
MAPE	3.48	7.68	2.53

Table 4.3: Metric result for BILSTM without attention

The performance comparison of the HYBRID with this model is plotted in bar graph shown on Figure4.7. This bar graph shows the all metric results that are obtained by the models in this case the HYBRID and the BILSTM without attention for the three sections of the training,validation and testing. Lower height having bar graph shows a model having lower metric value.

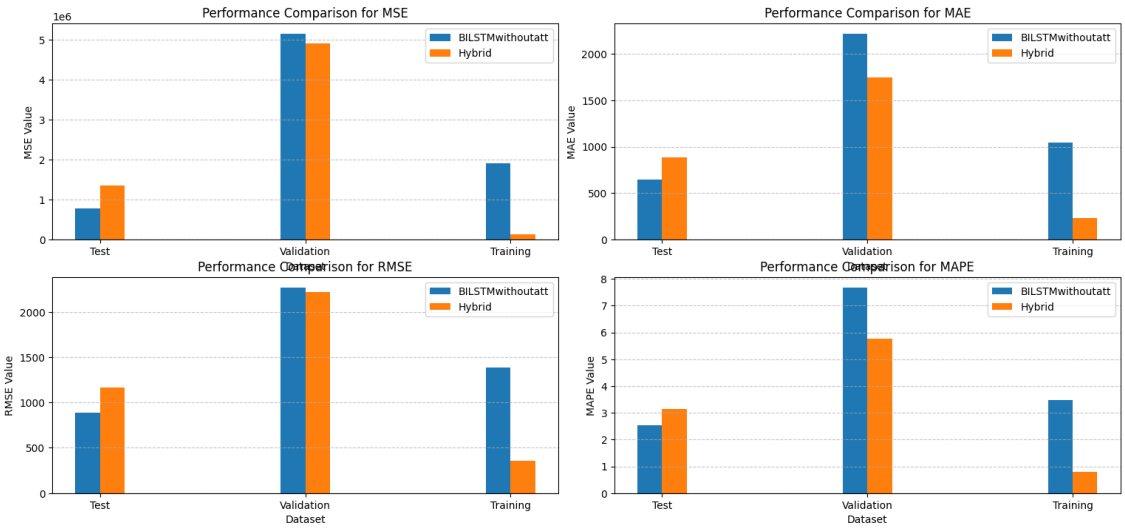

Figure 4.7: Performance bar graph of HYBRID vs BILSTM without Attention

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	93.22(Hybrid)	4.49(Hybrid)	72.54(BILSTMwithoutatt)
MAE	77.87(Hybrid)	21.36(Hybrid)	36.97 (BILSTMwithoutatt)
RMSE	73.97(Hybrid)	2.27(Hybrid)	31.35(BILSTMwithoutatt)
MAPE	77.30(Hybrid)	25.00(Hybrid)	23.72(BILSTMwithoutatt)

Table 4.4: Performance in percentage of HYBRID vs BILSTM without Attention

Table 4.4 shows percentage performance comparison where the HYBRID performed better initially it later on started to struggle during the validation period and ending to having a lower performance when coming to the test part of the data. We believe the attention mechanism being deducted from its structure has lead it to have these kind of results.BILSTM without attention model resulted in having lower error metric values in comparison to its counter part by having lower values of 72.54,36.97,31.35,23.72 percent on the test part of the data.

4.3.0.2 ARIMA

Alongside the above baseline experiments the other model used for performance comparison against the HYBRID model is the ARIMA model and their performance is shown in Figure 4.8.

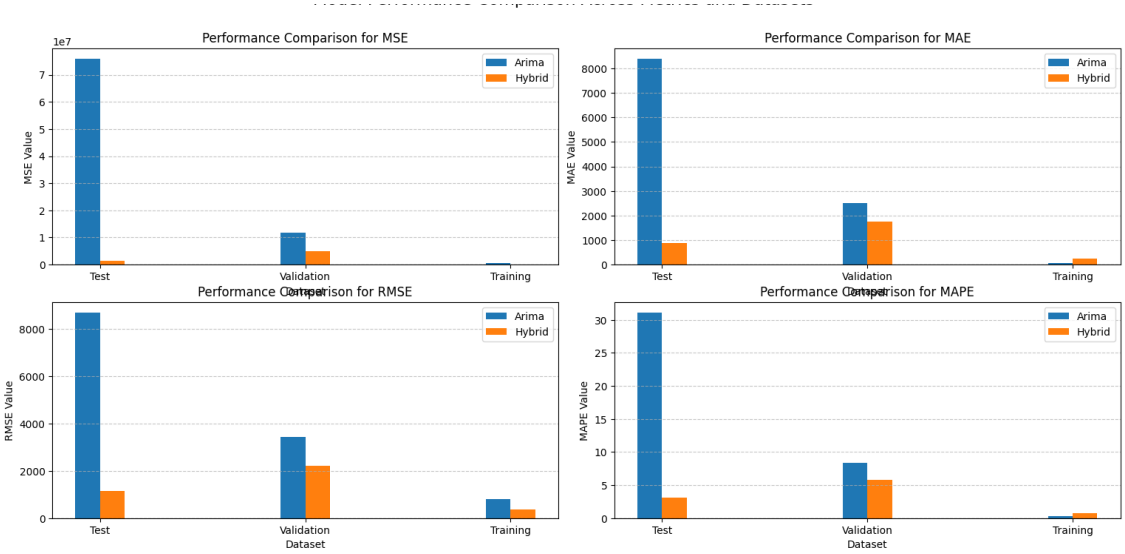

Figure 4.8: Performance graph of HYBRID vs ARIMA

From Figure 4.8 it can be seen that the Hybrid model outperforms the ARIMA model in all metric tests across all three sets (Training, Validation, and Test). This indicates that the Hybrid model is generally more accurate in predicting.This is further shown on Table 4.5 where the better performing model is shown in bracket.As it can be seen from the percentage increment that resulted from the ARIMA model it is an augmentation to the earlier results we had gotten when the prediction plot was done.Initially during the training part of the data the ARIMA performed better but during the validation the HYBRID model was beginning to outperform the ARIMA as it can be seen from the lower percentage performance results and finally in the testing part of the data the HYBRID outperformed the ARIMA in huge margins proving once again that the HYBRID is a better model in structure for dealing with the data at hand.

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	79.94(Hybrid)	58.19(Hybrid)	98.21(Hybrid)
MAE	194.27(ARIMA)	30.58(Hybrid)	89.43 (Hybrid)
RMSE	55.21(Hybrid)	35.34(Hybrid)	86.63(Hybrid)
MAPE	182.14(ARIMA)	31.26(Hybrid)	89.94(Hybrid)

Table 4.5: Performance in percentage of HYBRID vs ARIMA

HYBRID model resulted in having lower error metric values in comparison to its counter part by having lower values of 98.21,89.43,86.63,89.94 percent on the test part of the data.

4.4 Comparison against alternative models

4.4.0.1 BILSTM with attention

The first baseline model we took for this section of work is the BILSTM with attention model.The structure included the same selection process for the right attention mechanism and in similar manner the multi head attention got selected as the proper attention for the data.In addition the ARIMA prediction and residual outputs are not fed as its input. Figure 4.9 shows the prediction model for the BILSTM model with attention.The green dotted line is a demarcation between the training and the validation split data where as the purple dotted line is the separation of the validation and the test data.

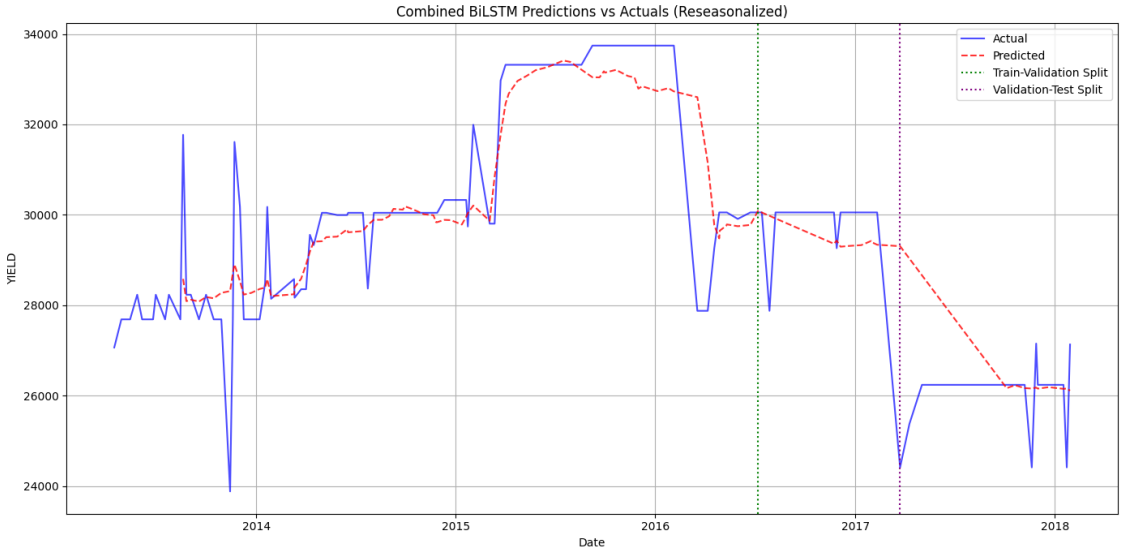

Figure 4.9: Prediction plot of BILSTM with attention Mechanism

From Figure 4.9 it can be seen that initially the model resulted in a prediction plot that almost resembles the actual data except for the final part of the testing data where it is falling short in mimicking the actual data.This prediction plot is achieved its best result at epoch of 200,batch size 32,RMSprop optimizer with validation loss of 0.3759,each L1 and L2 with 0.001 value. It achieved the following metric results along with the prediction plot which are shown on Table 4.6.Even though the model had multiple changes in its parameter this metric results and prediction plot were the ones that came close to resembling the actual data.

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	1222884.06	3045493.92	808705.02
MAE	667.86	1108.17	585.58
RMSE	1105.84	1745.13	899.28
MAPE	2.24	4.11	2.31

Table 4.6: BILSTM with attention metric result

In order to know by how much the HYBRID model improved or degraded the BILSTM model is seen in the comparison analysis we performed and based on the individual metric tests both got their comparison is shown in bar plot as well in percentage analysis.

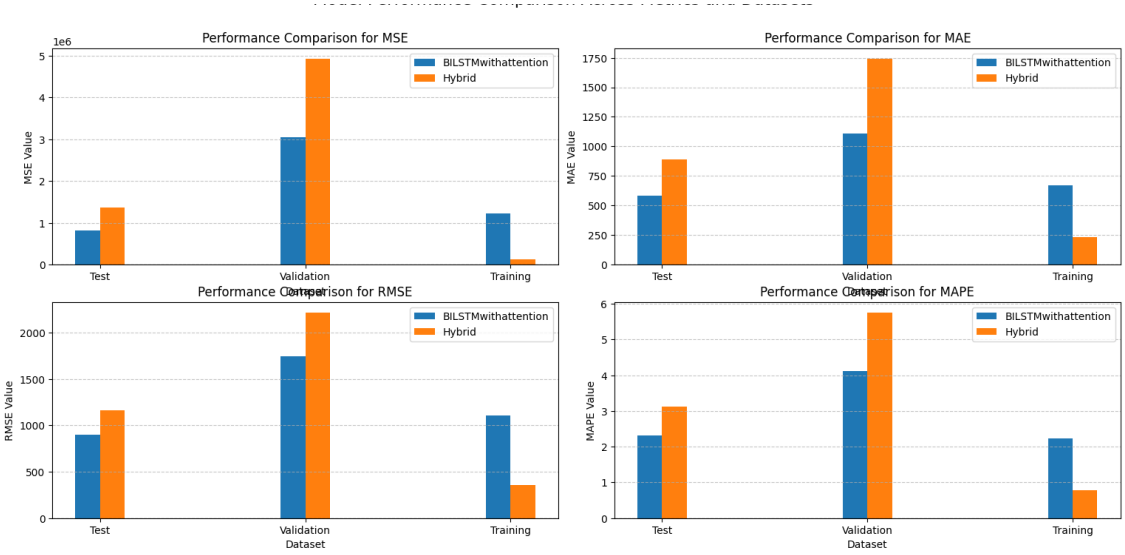

Figure 4.10: Performance comparison of BILSTM with attention Mechanism Vs HYBRID

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	89.04(HYBRID)	61.65(BILSTMwithatt)	67.75(BILSTMwithatt)
MAE	65.52(HYBRID)	57.09(BILSTMwithatt)	51.51(BILSTMwithatt)
RMSE	67.44(HYBRID)	27.14(BILSTMwithatt)	29.52(BILSTMwithatt)
MAPE	64.73(HYBRID)	40.15(BILSTMwithatt)	35.50(BILSTMwithatt)

Table 4.7: Performance of BILSTM with attention Vs HYBRID in percentage

As it can be seen on both the Figure 4.10 and on the Table 4.7 The BILSTM with attention is performing much better in reasons related to having the attention mechanism and lesser complex architecture. BILSTM with attention model resulted in having lower error metric values in comparison to its counter part by having lower values of 67.75 ,51.51,29.52,35.50 percent on the test part of the data.

4.4.1 Window Sliding

Seeing the above results have not yet met our expectation we decided to add another model which we believe can capture the actual data in a better way and for that we used Window Sliding model.This model is based on the thinking that by setting the window size,step size and the automatic failure point detection we can actually have a better performing model for doing the prediction work. The Prediction plot for this model is shown on Figure 4.11.

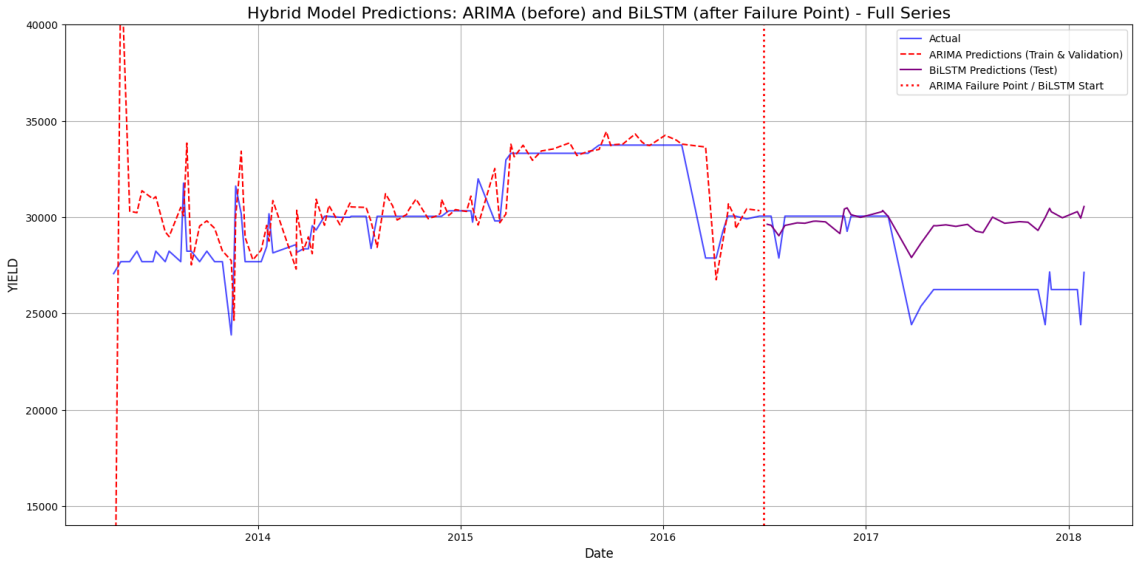

Figure 4.11: Sliding window model

As it can be seen from Figure 4.11 failure points of the ARIMA model are marked at middle of the year 2016 based on the automatic detection which uses RMSE result of the validation part of ARIMA as a reference. Any deterioration beyond double the value of RMSE for the validation value will be defined as a failure point. We used double the validation RMSE value so that the model does not become to sensitive if multiplier is set to low or become unresponsive to failure points unless they are numerically large enough if multiplier is higher than double of the validation metric value.We also added a rolling window to our model so as to make failure detection more robust and not to consider all spikes in the data as a error. We experimented our rolling window with the values of 5,7,10,14 and finally achieving a good result with a rolling window value of 5. We used this rolling window to do mean of the absolute errors over the last rolling window data points.After this the average of last rolling window data points is checked if it is above or below the already set threshold. Based on its result if checked to be higher than set baseline then BILSTM starts its work. The better result achieved from this model is also achieved by using L1 and L2 0.001, patience of 20, BILSTM units of 64, a look back of 10 in addition to the rolling window.The chosen optimizer for this model is RMSprop with a validation loss of 0.2174 The final metric results obtained after this experiment is shown in the following Table 4.8

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	919223.90	2329969.65	3491918.90
MAPE	2.52	3.90	4.99
RMSE	958.76	1526.42	1868.67
MAE	786.36	1032.82	1275.02

Table 4.8: Sliding Window model metric tests

Based on the results shown on Table 4.8 next we performed performance comparison with our proposed model and obtained the following results shown on Figure4.12 and on Table 4.9 shows in detail the performance of each model in percent.

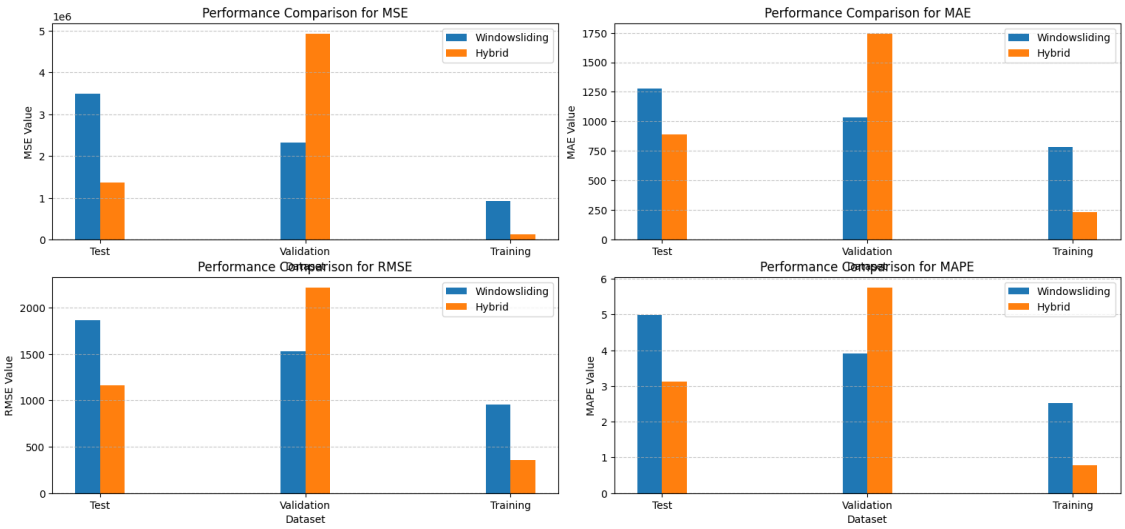

Figure 4.12: Sliding window model

Metric tests	Training Metrics	Validation Metrics	Test Metrics
MSE	85.89(HYBRID)	111.29(Window)	61.15(HYBRID)
MAPE	68.65(HYBRID)	47.69(Window)	37.27(HYBRID)
RMSE	62.44(HYBRID)	45.36(Window)	37.67(HYBRID)
MAE	70.72(HYBRID)	68.55(Window)	30.41(HYBRID)

Table 4.9: Sliding Window model Vs Hybrid model in percent

As it can be seen from the percentage result shown on Table 4.9 it can be seen that even though the HYBRID model is giving result of it performing better on the test part of the data it posses a slight confusion since the validation part of the data is shown to have the Window sliding model perform better.HYBRID model resulted in having lower error metric values in comparison to its counter part by having lower values of 61.15 ,37.27,37.67,30.41 percent on the test part of the data.

4.5 Discussion

Even though time series yield prediction is an area that is widely explored to our awareness this is the only paper that tries to explore HYBRID model building with the inclusion of both the prediction and residual of ARIMA model as an input, using a selection mechanism for selecting important features from both numerical and non numerical features,chose appropriate attention mechanism that soley depends on the structure of the model,Chose also the best optimizer that gives lower validation error and finally choosing appropriate attention mechanism suited for the work at hand.

To show the effectiveness of a HYBRID model that is only based on NDVI and weather data was the primary goal of the work and to do that an attention mechanism was incorporated to the proposed model and other models that are commonly used for time series prediction are also used to compare the improvement registered by the HYBRID model performance. The models that are used for comparison are ARIMA,BILSTM with attention,BILSTM without attention mechanism and Window Sliding model.

In answering our research question which were:-

1. RQ1:- What are the strengths and limitations of the hybrid ARIMA-BILSTM model in predicting wheat yield in comparison to traditional prediction models? The unpredicted data after failure of the ARIMA model is dealt well by the incorporation of the BILSTM model, The selection process for the applying an attention mechanism and its iteration through three optimizer for selecting the suitable one are among the strengths that are found from this work while its limitation is the complex structure that is created when both prediction and residual of ARIMA along with other features fed as an input to the BILSTM model.As a result it struggled to perform from the limited amount of the dataset. Especially not having factors such as soil,fertilizer and other affecting factors as part of our dataset which in turn has created a problem of having a limited training data which led our model to act as an overfit for the rest of the data which was the validation and test parts. In addition to the increment of dataset features an increment in period that is beyond the amount of years included in our dataset we believe can also have a different outcome than the ones shown during our experimentation period.

2. RQ2:- How does the attention mechanism amplify the prediction achieved by the HYBRID model? The application of the attention mechanism has added complexity to our model instead of helping the prediction work which is getting done at a greater computational cost.

3. RQ3:- What other model kinds can be discussed to address the performance gap that will be seen if Hybrid model is not effective in its performance? The sliding window approach is one model in our belief has great potential in achieving great results, especially it shows in great detail the failure point obtained by ARIMA and despite of it having conflicting results during comparison with the HYBRID model we believe that if structure changing and if further tuning of parameters is performed it will be an easier model in comparison to the proposed one in understanding its working mechanism which we believe is beneficial for any researcher taking an interest in it.

4. RQ4:- Which environmental features which are highly impacting the Yield? From the feature importance that is based on the correlation matrix ,normalized difference vegetative index,minimum temperature and solar radiation are shown as the most highly impacting variables followed by maximum temperature, vapor pressure, date of measurement and rain snow fall.

Table 4.10 shows all models with the experiments that are done on and from it the following noticeable points can be made from it

- BILSTM with attention uses lower number of LSTM units in its structure in comparison to the other models
- depending on the level of complexity the models selected expected optimizers for their work
- Despite the different number of units used in their structure,regularizer values or their optimizer selection it can be seen that the Window Sliding has the lowest validation loss making it the better model for the validation part of the dataset.
- The addition of an attention mechanism has helped the BILSTM with attention model to have lesser number of units in its structure in comparison to its counter part the BILSTM without the attention creating a lesser complex model.

Models	no of units	Validation loss	Optimizer	L1,L2	Patience
HYBRID	100	3.7475	ADAM	0.1,0.1	70
BILSTM With att.	200	0.3759	RMSprop	0.001,0.001	20
BILSTM Without att.	32	0.2868	SGD	0.001,0.001	20
Window Sliding	200	0.1413	ADAM	0.001,0.001	20

Table 4.10: Comparison of Models

Chapter 5

Conclusion and Recommendation

Many findings have been made on this work of wheat yield prediction based on dataset containing weather and yield data. The dataset included these elements and not soil,fertilizer and other affecting factors.The HYBRID model showed an excellent performance on all metric results by having a lower result in comparison to metrics of ARIMA when it resulted in 98.21 percent on MSE ,a 89.43 percent on MAE, a 86.63 percent on RMSE and a 89.94 percent on the MAPE lower in comparison to its counterpart. In contrary the BILSTM model without attention mechanism which showed lower metric results than the HYBRID when it resulted in 72.54 percent on MSE,a 36.97 percent on MAE, a 31.35 percent on RMSE and a 23.72 percent on MAPE.And When compared to the alternative models of BILSTM with attention which scored lower results on all metrics tests by having of a 67.75 percent on MSE, a 51.51 percent on MAE,a 29.52 percent on RMSE and a 35.50 percent on MAPE.During performance comparison of the HYBRID model against the Window Sliding model it was found that the HYBRID achieved a lower error value on all the metrics by having a 61.15 percent on MSE, a 37.27 percent on MAPE,a 37.67 percent on RMSE and a 30.41 percent on MAE .In addition we also found the failure point for the ARIMA model is at the middle of the year of 2016 for the Window Sliding model.

The results obtained showcased the importance of HYBRID approaches for dealing with datasets having both linear and non linear parts.And architecture choice for our model in the fact that applying the attention mechanism in addition to the ARIMA output being used as an input has made the structure complex enough that its performance degraded as a result. The experiments we did highlight this issue due to the other models not using any of the ARIMA output has resulted them in having a better performance when compared to our proposed model. In conclusion the HYBRID model technique has shown great findings in relation to architectural choice for this work of yield prediction through its results that showcase that dealing with complex and nonlinear parts of the data is a crucial activity in yield prediction. And from our work we believe that we can recommend future works targeting yield prediction through HYBRID modeling architecture should be deeply investigate by researchers so as not to have unrealistic model which fails in understanding the work at hand.

Furthermore, efforts must be focused on creating a model that can take these findings and observations and do yield prediction work. The possibility of building a more productive HYBRID model for yield prediction in the long run depends on doing further investigation on the combination of attention mechanism and structure of model.

Bibliography

[1] Nicholas Minot, James Warner, Solomon Lemma, Leulsegged Kasa, Abate Gashaw, and Shahidur Rashid. The wheat supply chain in ethiopia: Patterns, trends, and policy options. *Gates Open Res*, 3(174):174, 2019.

[2] Sándor Zsebő, László Bede, Gábor Kukorelli, István Mihály Kulmány, Gábor Milics, Dávid Stencinger, Gergely Teschner, Zoltán Varga, Viktória Vona, and Attila József Kovács. Yield prediction using ndvi values from greenseeker and micasense cameras at different stages of winter wheat phenology. *Drones*, 8(3):88, 2024.

[3] Team Cropin. Ndvi and its practical applications in agriculture. *Drones*, 2024.

[4] N.Sri A. Saikrishna T. Rajinikanth, B.Kavya. Agriculture crop yield analysis and prediction using feature selection based machine learning techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, 12, 2022.

[5] Amal Mahmoud, Ammar Mohammed, AA Khalil, et al. Time series forecasting of wheat crop productivity in egypt using deep learning techniques. *International Journal of Data Science and Analytics*, pages 1–16, 2024.

[6] Benjamin Kwapong Osibo, Tinghuai Ma, Mohamed Magdy Abdel Wahab, Li Jia, Ye Wenzheng, Bright Bediako-Kyeremeh, and Stephen Osei-Appiah. Interpretable deep learning model for crop yield prediction: A case study of wheat yield prediction in egypt. *Research Square*, 2023.

[7] Yong Shi, Jianyu Miao, Zhengyu Wang, Peng Zhang, and Lingfeng Niu. Feature selection with $\ell_{2,1-2}$ regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4967–4982, 2018.

[8] Chao Fan, Pei-Ge Cao, Tie-Jun Yang, and Hong-Liang Fu. Research on the prediction model of grain yield based on the arima method. In *2015 4th International Conference on Sensors, Measurement and Intelligent Materials*, pages 454–458. Atlantis Press, 2016.

[9] Sowmya and Prasad. Integrated approach for crop yield prediction in telangana region using ensemble techniques and arima model. *Available at SSRN 4762412*, 2024.

[10] S Kannan and KM Karuppasamy. Forecasting for agricultural production usisng arima model. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 17(9):5939–49, 2020.

[11] Santosha Rathod, KN Singh, Prawin Arya, Mrinmoy Ray, Anirban Mukherjee, Kanchan Sinha, Prakash Kumar, and Ravindra Singh Shekhawat. Forecasting maize yield using arima-genetic algorithm approach. *Outlook on Agriculture*, 46(4):265–271, 2017.

[12] Wanhyun Cho, Sangkyuoon Kim, Myunghwan Na, and Inseop Na. Forecasting of tomato yields using attention-based lstm network and arma model. *Electronics*, 10(13):1576, 2021.

[13] Ligang Cui, Yingcong Chen, Jie Deng, and Zhiyuan Han. A novel attlstm framework combining the attention mechanism and bidirectional lstm for demand forecasting. *Expert Systems with Applications*, page 124409, 2024.

[14] Hossein Abbasimehr and Reza Paki. Improving time series forecasting using lstm and attention models. *Journal of Ambient Intelligence and Humanized Computing*, 13(1):673–691, 2022.

[15] Luocheng Liang. Arima with attention-based cnn-lstm and xgboost hybrid model for stock prediction in the us stock market. In *SHS Web of Conferences*, volume 196, page 02001. EDP Sciences, 2024.

[16] Alexandros Oikonomidis, Cagatay Catal, and Ayalew Kassahun. Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, 36(1):2031822, 2022.

[17] Sonal Agarwal and Sandhya Tarar. A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. In *Journal of Physics: Conference Series*, volume 1714, page 012012. IOP Publishing, 2021.

[18] Pan Wu, Zilin Huang, Yuzhuang Pian, Lunhui Xu, Jinlong Li, and Kaixun Chen. A combined deep learning method with attention-based lstm model for short-term traffic speed forecasting. *Journal of Advanced Transportation*, 2020(1):8863724, 2020.

[19] Nandini Geddlehally Renukaradya, Kishore Gopala Rao, and Anand Babu Jayachandra. Classification and estimation of crop yield prediction in karnataka using lstm with attention mechanism. *International Journal of Intelligent Systems and Applications in Engineering*, 12(3):89–96, 2024.

[20] Kun Zhou, Wen Yong Wang, Teng Hu, and Chen Huang Wu. Comparison of time series forecasting based on statistical arima model and lstm with attention mechanism. In *Journal of physics: conference series*, volume 1631, page 012141. IOP Publishing, 2020.

[21] Fan Liu, Xiangtao Jiang, and Zhenyu Wu. Attention mechanism-combined lstm for grain yield prediction in china using multi-source satellite imagery. *Sustainability*, 15(12):9210, 2023.

[22] Wei Xiang, Long Long, Zichen Liu, Feng Dai, Yucheng Zhang, Hu Li, and Lin Cheng. A crop model based on dual attention mechanism for large area adaptive yield prediction. *Available at SSRN 5023172*, 2024.

[23] Johnathon Shook, Tryambak Gangopadhyay, Linjiang Wu, Baskar Ganapathysubramanian, Soumik Sarkar, and Asheesh K Singh. Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, 16(6):e0252402, 2021.

[24] Huiren Tian, Pengxin Wang, Kevin Tansey, Dong Han, Jingqi Zhang, Shuyu Zhang, and Hongmei Li. A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the guanzhong plain, pr china. *International Journal of Applied Earth Observation and Geoinformation*, 102:102375, 2021.

[25] Sunday Samuel Olofintuyi, Emmanuel Ajayi Olajubu, and Deji Olanike. An ensemble deep learning approach for predicting cocoa yield. *Heliyon*, 9(4), 2023.

[26] Sajjad Ali Haider, Syed Rameez Naqvi, Tallha Akram, Gulfam Ahmad Umar, Aamir Shahzad, Muhammad Rafiq Sial, Shoaib Khaliq, and Muhammad Kamran. Lstm neural network based forecasting model for wheat production in pakistan. *Agronomy*, 9(2):72, 2019.

[27] PRITESH PATIL, PRANAV ATHAVALE, MANAS BOTHARA, SIDDHI TAMBOLKAR, and ADITYA MORE. Crop selection and yield prediction using machine learning approach. *Current Agriculture Research Journal*, 11(3), 2023.

[28] Manisha Galphade, Nilkamal More, Abhishek Wagh, and VB Nikam. Crop yield prediction using weather data and ndvi time series data. In *Advances in Data Computing, Communication and Security: Proceedings of I3CS2021*, pages 261–271. Springer, 2022.

[29] André Barriguinha, Bruno Jardim, Miguel de Castro Neto, and Artur Gil. Using ndvi, climate data and machine learning to estimate yield in the douro wine region. *International Journal of Applied Earth Observation and Geoinformation*, 114:103069, 2022.

[30] Khulood Albeladi, Bassam Zafar, and Ahmed Mueen. Time series forecasting using lstm and arima. *International Journal of Advanced Computer Science and Applications*, 14(1):313–320, 2023.

[31] K Pravallika, G Karuna, K Anuradha, and V Srilakshmi. Deep neural network model for proficient crop yield prediction. In *E3S web of conferences*, volume 309, page 01031. EDP Sciences, 2021.

[32] Ghahreman Abdoli. Comparing the prediction accuracy of lstm and arima models for time-series with permanent fluctuation. *Periódico do Núcleo de Estudos e Pesquisas sobre Gênero e DireitovCentro de Ciências Jurídicas-Universidade Federal da Paraíba*, 9, 2020.

[33] Halit Çetiner and Burhan Kara. Recurrent neural network based model development for wheat yield forecasting. *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, 9(16):204–218, 2022.

[34] Haoming Mo, Ying Zhang, Yifei Liu, and Yanzi Zheng. Prediction of rice yield based on lstm long short term memory network. In *Journal of Physics: Conference Series*, volume 1952, page 042033. IOP Publishing, 2021.

[35] Kavita Jhajharia, Pratistha Mathur, Sanchit Jain, and Sukriti Nijhawan. Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, 218:406–417, 2023.

[36] Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Moez Krichen, and Charles Lebon Mberi Kimpolo. Crops yield prediction based on machine learning models: Case of west african countries. *Smart Agricultural Technology*, 2:100049, 2022.

[37] Alejandro Morales and Francisco J Villalobos. Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, 14:1128388, 2023.

[38] D Jayanarayana Reddy and M Rudra Kumar. Crop yield prediction using machine learning algorithm. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1466–1470. IEEE, 2021.

[39] Rajswee Surana and Ritu Khandelwal. Crop yield prediction using machine learning: A pragmatic approach. *Research Square*, 2024.

[40] Jing Lin Ng, Yuk Feng Huang, Stephen Luo Sheng Yong, Jin Chai Lee, Ali Najah Ahmed, and Majid Mirzaei. Analysing the variability of non-stationary extreme rainfall events amidst climate change in east malaysia. *AQUA - Water Infrastructure, Ecosystems and Society*, 73(7):1494–1509, 07 2024.