IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 2, Ser. 3 (Mar. – Apr. 2024), PP 23-32
www.iosrjournals.org

Resilient Machine Learning Model For Seasonal Rainfall Forecasting Using Enhanced Forecast Data

Roy Chaikatisha[1], Mayumbo Nyirenda[2]
(Department Of Computer Science, University Of Zambia, Zambia)

Abstract:
The hype surrounding artificial intelligence and data sciences has prompted renewed interest in the application of machine learning techniques to the field of meteorology. Contemporary research has led to the exploration of neural networks as candidates for weather prediction when combined with massive data sets. Such research has shown that artificial neural networks can successfully be applied to prediction of seasonal rainfall. However, it is evident that the accuracy of these models is compromised by poor quality data emanating from gaps in historical data as well as inconsistent and irregular collection of historical data. Quality of data available is therefore a factor that impacts the skill of the forecasts. Given the foregoing limitations, there is increased interest in the study of machine learning models that are resilient to poor data quality as well as techniques that can be used to enhance data quality.
This study therefore sought to contribute further to the enhancement of seasonal rainfall forecasting in Zambia. Particularly, the study sought to explore machine learning models that are resilient to data gaps as well as approaches that can be used to resolve the data gaps while not compromising the efficacy of the seasonal rainfall forecasts. The study utilized Zambian rainfall data collected from the Zambia Meteorology Department. This comprised rainfall data from 1981 to 2020. Several experiments were carried out to inject faults into the dataset in order to ascertain which models are resilient to poor quality data and data gaps. The results showed that models exist that are resilient to poor quality data and data gaps within reasonable bounds. The study further explored the availability of external datasets that can be used to fill in missing data for rainfall forecasting. Using literature review and data experiments, the study determined that datasets such as reanalysis and Enact data, which combine satellite and station data, have reliable historical data spanning beyond 30 years. This can thus be a good source of historical weather data that can be used to fill in gaps and enhance rainfall forecasts in Zambia.

Keywords: *Zambia Meteorology Department, Fault Injection, Reanalysis, Enact, Data quality, Artificial Intelligence, Machine Learning.*

Date Of Submission: 04-04-2024 Date Of Acceptance: 14-04-2024

I. Introduction

The use of artificial intelligence and machine learning in weather prediction has gained popularity, with a particular focus on improving rainfall forecasts through neural networks and large datasets. However, the accuracy of these forecasts is often hindered by the poor quality of historical data. This is characterized by data gaps and inconsistencies in collection methods. To overcome these challenges, there's a growing need to explore the use of resilient machine models which are capable of handling data quality issues and addressing missing data without compromising forecast accuracy.

Appreciating that these challenges are pivotal in the advancement of contemporary seasonal rainfall prediction techniques, this study embarks on a thorough exploration aimed at refining seasonal rainfall forecasting practices in the Zambian context. Among the keys to this endeavor is the identification of resilient machine learning models whose performance holds up well varying data quality issues. Furthermore, the study endeavors to devise practical frameworks designed to address existing data gaps while not compromising the accuracy of forecast outcomes.

II. Related Works

Prior research has extensively investigated the application of machine learning techniques in the domain of weather prediction.[1] For example, Artificial Neural Networks (ANNs) have been used with great success in seasonal rainfall forecasting due to their ability to handle non-linear relationships and provide more accurate predictions. [2] Besides weather prediction, ANNs have also been successfully employed in other disciplines, highlighting their varied and extensive use in artificial intelligence. [3] Other studies have focused on exploring

machine leaning models' resilience when faced with poor quality data and data gaps. Additionally, recent research has sought to explore enhancing forecast data through the incorporation of external datasets. This section reviews relevant studies in this field, examining advancements in machine learning based rainfall forecasting and identifying areas for further research and refinement.

Mzyece et al [4] shed light on the limitations of conventional statistical models and regression analysis in accurately predicting seasonal rainfall. These traditional methods often fail to consider various influencing factors, leading to diminishing accuracy over time. In contrast, the study underscores the efficacy of ANNs in rainfall forecasting. [5] While Mzyece et al [4] highlight the superiority of ANNs over traditional methods, there is a lack of analysis to investigate and compare the performance of ANNs against alternative machine learning algorithms within the context of seasonal rainfall forecasting.

Abraham et al. [6], while not addressing seasonal rainfall forecasting, contributed to understanding the resilience of neural network ensembles against faulty training data. Their study emphasized the effectiveness of ensemble learning, wherein results from multiple machine learning models are combined, in exhibiting greater resilience compared to individual models when faced with faulty training data. The findings underscored that a model's performance on clean training data does not necessarily translate to performance when the training data is faulty, highlighting the importance of assessing model robustness under diverse conditions. However, potential gaps exist due to the study's reliance on specific metrics for resilience evaluation, such as the Gini coefficient and the Shannon equitability index, which warrant exploration of additional metrics to provide a more comprehensive assessment of ensemble performance.

In their study, Sumi et al [7] conducted a comprehensive comparative analysis of various machine learning models for rainfall forecasting, encompassing ANNs, Support Vector Machines (SVM), Multivariate Adaptive Regression Splines (MARS), and k-Nearest Neighbors (kNN). The research aimed to assess the efficacy of these models in predicting rainfall, particularly focusing on the case of Fukuoka city. One notable contribution of the study is the proposal of a hybrid multi-model approach for rainfall forecasting, which integrates the strengths of individual models to enhance overall forecasting accuracy. However, a possible limitation of the study is the lack of thorough discussion regarding the quality of the rainfall data utilized in the forecasting models. Issues such as missing values, measurement errors, or inconsistencies in the data were not adequately addressed, potentially impacting the reliability and accuracy of the forecasting models.

A study by Ferrari, Ozaki, et al [8] focused on filling missing rainfall data in the Luvuvhu River Catchment in South Africa, using ANNs. The research aimed to address the challenge of incomplete precipitation data by comparing different imputation methods. The findings revealed that the inverse distance weighting method emerged as the most effective imputation method for precipitation data, demonstrating the lowest mean absolute error and root mean square error. However, a notable limitation of the study was the lack of validation of imputed data with independent observations or other methods. This absence of validation could potentially impact the reliability and accuracy of the results.

Holmstrom et al [9] similarly explored the application of machine learning in weather forecasting. Their study focused on comparing the performance of linear regression and functional regression models. The findings revealed that linear regression outperformed functional regression for all forecasts. However, a limitation of the study was the exclusive use of linear regression and functional regression models.

This study sought to glean from contemporary research in this field, examining advancements in machine learning based rainfall forecasting approaches and identifying research gaps to derive areas for further research and refinement.

III. Material And Methods

Research Design

A mixed-methods research design was employed to achieve the objectives of this study. This aimed to explore machine learning models' resilience to data gaps and identifying approaches to address these gaps without compromising seasonal rainfall forecast efficacy. The CRISP-DM (Cross Industry Standard Process for Data Mining) was used to achieve the data science objectives of the study. Whereas a conceptual model was devised to guide the experiments of the study.

CRISP-DM Model for Data Science Exploration

The CRISP-DM provides an overview of the life cycle of a data science project. [10] This methodology was used owing to its suitability for data science-oriented research. The CRISP-DM has the following sequential phases, Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. [11] These were employed to achieve the study's objectives.

Business understanding

This involves understanding the data science requirements. Devising a data mining aim is one of the most crucial aspects of this stage. The type of data mining to be conducted, such as classification or regression analysis should be understood first. [12]. In this vein, this study sought to perform a classification data mining. This is line with current practice at the Zambia Meteorological department. The goal is to classify the forecasted rainfall as either below normal, normal or above normal rainfall. According to a Zambian researcher, in the existing procedure, precipitation data spanning from 1981 to the present is arranged in ascending order using Microsoft Excel. Subsequently, the data is visually divided into three segments through a subjective inspection. Following this, the projected rainfall value is assessed by visually comparing it to the sorted data, placing the forecasted figure in proximity to the corresponding value to categorize it as below normal, normal, or above normal rainfall. [4] One of the goals of the research is to replace the visual inspection aspects with automated techniques. The proposed approach will borrow the key concepts as carried out by the ZMD but ensure an automated approach.

Data understanding

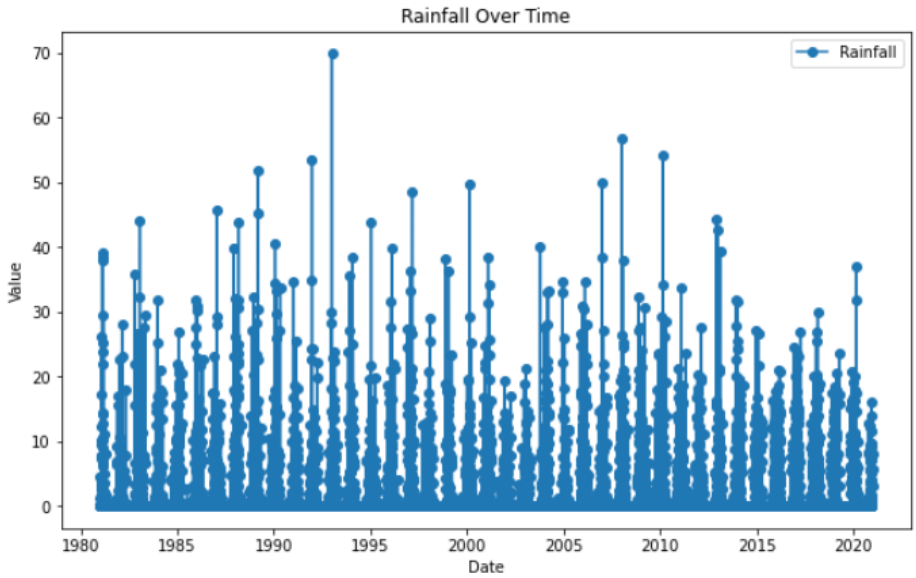
This stage involved collecting required data, exploring it to gain key insights, and assessing the quality of data. [13] In the context of this research, primary data was obtained from the ZMD, spanning from 1981 to 2020. The dataset had 14610 records with the following features as seen in table 1 below.

Table 1. Selected Features

Feature	Description
Date	The date feature indicates the specific calendar day when weather measurements were recorded.
Max Temperature	The maximum temperature feature represents the highest recorded air temperature during a specified period.
Min Temperature	The minimum temperature feature denotes the lowest recorded air temperature within a specific time frame.
Rainfall	The rainfall feature quantifies the amount of precipitation, usually in millimeters or inches, that has occurred over a specific timeframe.

Training data was then plotted to confirm seasonality trend. Seasonal patterns are often seen as recurring trends or cycles within the graphical representation. [14] This is seen in figure 1 below, thus confirming the periodic and seasonal nature of the dataset which was selected. These patterns typically exhibit regular and predictable fluctuations, highlighting variations that repeat over specific intervals. When observing a seasonal plot, you might identify peaks and troughs, illustrating the periodic nature of the data.

Figure 1. Plotted weather dataset showing seasonality



Data Preparation

This stage involved steps such as, data normalization, transformation, and statistical binning. [15]. Temperature data was converted from tenths of degrees Celsius to the standard degrees Celsius format. This was

done by dividing the value for each temperature feature by 10. The data was transformed according to the algorithm below.

1. Aggregate Daily Rainfall, Maximum temperature, Minimum temperature and group by year and month.
2. Split aggregated data into 12 datasets, one for each month.
3. For each data set representing a month, classify data into bin quartiles representing Below average, Average and Above average rainfall.
4. Merge 12 datasets back into a single data set.

This transformation therefore successfully automated the corresponding steps of the current ZMD procedure. Figure 2 below shows how datasets for each month were labelled as Below average, Average and Above average. A new feature, 'Total Rain Quantile' was then created to hold the rainfall categories.

Figure 2. Statistical binning

Bins represent , below average, average, above average

```
In [19]: # Apply binning for all the dataframes
df1['Total Rain Quantile'] = pd.cut(df1['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df2['Total Rain Quantile'] = pd.cut(df2['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df3['Total Rain Quantile'] = pd.cut(df3['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df4['Total Rain Quantile'] = pd.cut(df4['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df5['Total Rain Quantile'] = pd.cut(df5['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df6['Total Rain Quantile'] = pd.cut(df6['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df7['Total Rain Quantile'] = pd.cut(df7['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df8['Total Rain Quantile'] = pd.cut(df8['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df9['Total Rain Quantile'] = pd.cut(df9['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'], in
df10['Total Rain Quantile'] = pd.cut(df10['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'],
df11['Total Rain Quantile'] = pd.cut(df11['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'],
df12['Total Rain Quantile'] = pd.cut(df12['Total Rain'], bins=3, labels = ['A-Below Average','B-Average','C-Above Average'],
```

The three categories are ordinal categories as they have a specified order. Label encoding was done to transform labels into numerical values that can be consumed by machine learning estimators. Figure 3 below shows the resulting dataset with labelled data.

Figure 3. Sample classified and labelled data

```
[34]: merged_weather_df.tail(5)
```

Out[34]:

	Date	Max Temp	Min Temp	Total Rain	Total Rain Quantile	rainfall_label	Year	Month
475	2016-12-01	35.7	15.7	152.677	B-Average	1	2016	12
476	2017-12-01	36.2	17.3	105.362	B-Average	1	2017	12
477	2018-12-01	38.0	17.1	99.552	B-Average	1	2018	12
478	2019-12-01	38.8	16.5	100.475	B-Average	1	2019	12
479	2020-12-01	36.7	15.6	133.018	B-Average	1	2020	12

Modelling

The study employed a diverse set of machine learning models, including Random Forests [16], XGBoost [17], and Logistic Regression [18] which are known for their adaptability to varying data conditions such as missing and faulty data. [13]. The dataset underwent a distinctive partition into training and testing sets, deviating from the conventional 80%–20% ratio. In consideration of the seasonal nature inherent in the data, the training phase incorporated inputs spanning from the year 1981 to 2019. Subsequently, the year 2020 was exclusively designated for the testing set. This tailored approach acknowledges the chronological progression of the data and ensures a robust evaluation of the models, enhancing their ability to generalize and make accurate predictions on unseen, future data.

Evaluation

The evaluation of the models encompassed the utilization of key performance metrics such as accuracy, precision, recall, and F1-score. [19] These metrics provided a comprehensive insight into the models' proficiency in managing data gaps while maintaining the precision of seasonal rainfall forecasts. Additionally, a thorough comparative analysis was conducted, comparing the outcomes derived from various machine learning models. This comparative examination aimed to discern the specific strengths and weaknesses exhibited by each model when confronted with data gaps. The findings of this comparative analysis play a key role in enhancing the overall reliability of seasonal rainfall predictions.

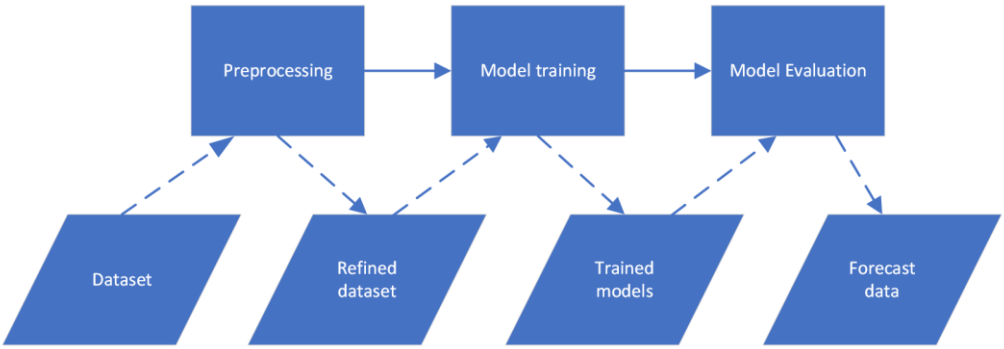
Conceptual Model

A framework was devised to assist guide the experiments that were carried out. This was done with a view to determine the selected models' performance against data gaps and faults. Firstly, collected data was first trained without faults. All selected models were used on the same dataset with relevant hyperparameters set accordingly. Secondly, faults were introduced to the training data then all the models were applied to the same dataset. The results from first, second and other experiments were then compared to determine which models are resilient to faulty data. As per information provided by the Zambia Meteorology Department (ZMD), a predominant challenge encountered in dealing with faulty data is the issue of missing data. [4] To thoroughly assess the effectiveness of machine learning models in addressing this challenge, the study aimed to formulate a mechanism for deliberately introducing missing data into the datasets. This deliberate introduction of missing data serves as a simulated test scenario, allowing for a comprehensive evaluation of how well the selected models can handle and adapt to the challenges posed by data gaps.

Model training without fault injection

The initial experiments were carried out without faults as illustrated in the figure 4 below. The initial phase of experimentation involved training the models on a dataset without faults. Specifically, the selected machine learning models, including Random Forests, XGBoost, and MLPClassifier, were employed on the original dataset. This preliminary step aimed to establish a baseline performance and comprehension of how the models respond to the input data under normal, fault-free conditions. By initiating the experiments without introducing faults, the study sought to evaluate the baseline capabilities and predictive accuracy of these models in handling the unaltered dataset.

Figure 4. Model training without fault injection

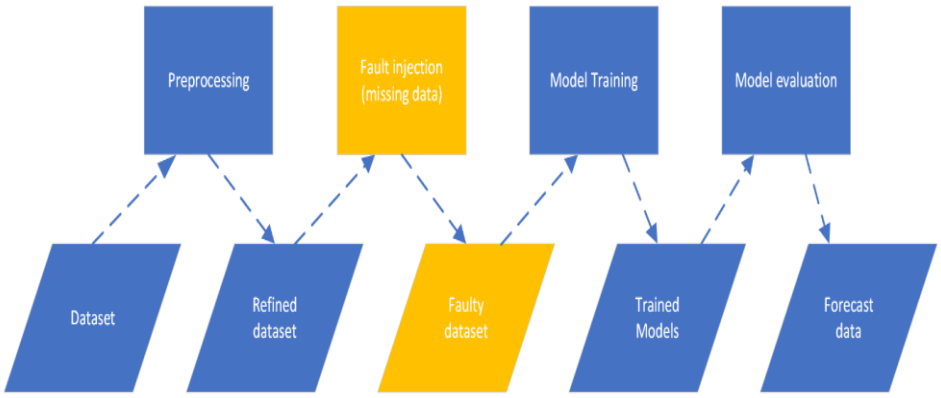


Model training with fault injection

In subsequent experiments, deliberate faults, specifically in the form of missing data, were introduced to the dataset. This intentional introduction of faults sought to emulate real-world challenges associated with data gaps, a critical issue acknowledged by the Zambia Meteorology Department (ZMD) [4]. During this phase, the machine learning models—Random Forests, XGBoost, and MLPClassifier and Logistic Regression—were rigorously evaluated to assess their resilience in addressing the prevalent challenge of missing data.

Figure 5 illustrates how the experiments were carried out at this stage.

Figure 5. Model training with fault injection



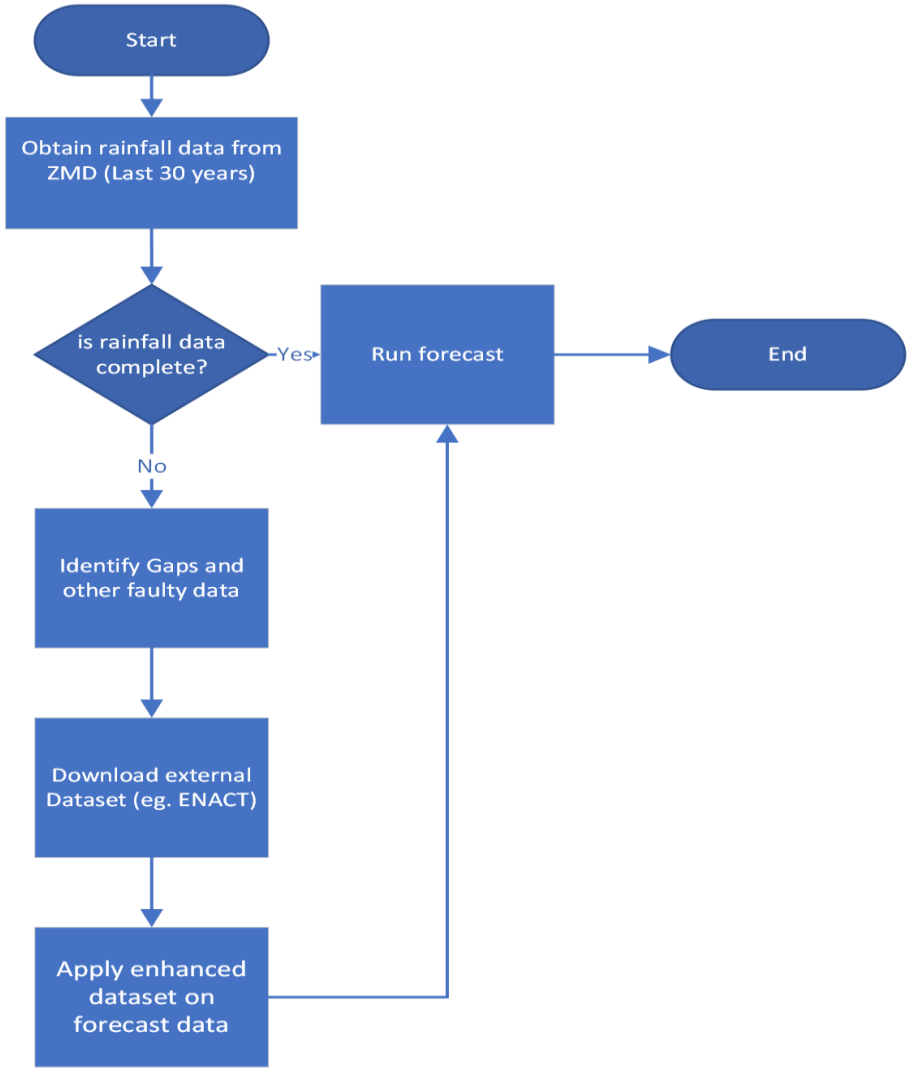
Framework for external datasets

With the aim of enhancing data quality through the utilization of external datasets, the study developed a framework designed to integrate external datasets for the purpose of filling missing data gaps and enhancing the accuracy of forecasts. By leveraging these external datasets, such as reanalysis and Enact data that amalgamate satellite and station data, the study aimed to reinforce the integrity of historical weather data. [20] This innovative approach sought to fortify the predictive capabilities of machine learning models, particularly in the context of addressing the persistent challenge of missing data, thereby contributing to more robust and reliable seasonal rainfall forecasts.

The framework defines an algorithm which begins by acquiring rainfall data spanning the last 30 years from the Zambia Meteorology Department (ZMD). It initiates a conditional check to determine the completeness of the rainfall data. If the data is complete, the algorithm proceeds to execute the forecast. In the event of incomplete data, the algorithm identifies the gaps. The external datasets are then downloaded. Subsequently, it applies the enhanced dataset to the forecast data, effectively enhancing forecast data via supplementary information. Finally, the algorithm executes the forecast process, incorporating both the original and enhanced datasets to enhance the accuracy and reliability of the predictions.

Figure 6 outlines the framework.

Figure 6. Framework for incorporating external datasets



Building a Prototype for incorporating external data

Following the development of a framework to leverage external datasets, the study constructed a prototype which implemented the framework. The prototype was developed using the Python and would be run from the command prompt. This prototype serves as a practical implementation of the framework designed to

enhance data quality by integrating external datasets. Aligned with this objective, the prototype is equipped to handle the key aspects of the framework. In its operation, the Python command prompt prompts the user to input the filename of the original dataset and subsequently requests the filename of the external dataset, which should be downloaded in advance. Upon execution, the prototype conducts a thorough analysis of the original dataset, checking for gaps. If no gaps are detected, it provides feedback indicating the absence of gaps. Conversely, if gaps are identified, the prototype dynamically fills these gaps and offers feedback acknowledging the successful completion of the gap-filling process. The prototype concludes by notifying the user that the forecast can now be executed seamlessly, showcasing the robustness of the model against poor quality data.

IV. Result and Discussion

This section explores the performance outcomes achieved by the models deployed in our study. The analysis encompasses a detailed examination of key metrics, including accuracy, precision, recall, and F1-score, offering an understanding of each model's effectiveness in handling the workings of seasonal rainfall forecasting.

Performance Analysis – without faults

In line with the conceptual model, the initial phase of experimentation involved training the models on a dataset without faults. The selected machine learning models, including Random Forests, XGBoost, Logistic Regression and MLPClassifier, were employed on the original dataset. This was done to have a baseline performance assessment, before faults are introduced into the dataset.

Random Forests

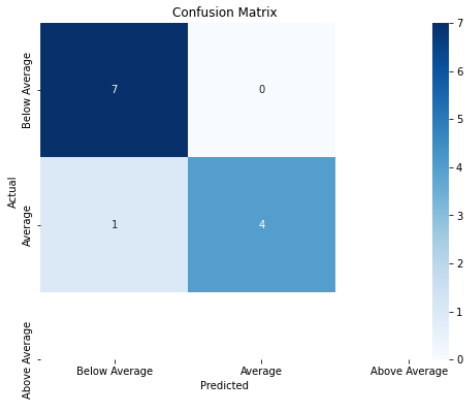
After the initial round of experiments, the Classification report for Random Forest model exhibited a high score in precision, recall and f1-score, with about 91% overall accuracy. Figure 7 below shows the classification report sample.

Figure 7. Classification report for Random Forest, first experiment

```
Accuracy = 0.9166666666666666
Cohen's Kappa = 0.8235294117647058
Time taken = 0.24726104736328125
              precision    recall  f1-score   support

           0    0.87500   1.00000   0.93333         7
           1    1.00000   0.80000   0.88889         5

    accuracy                        0.91667        12
   macro avg    0.93750   0.90000   0.91111        12
weighted avg    0.92708   0.91667   0.91481        12
```

Figure 8 below shows the confusion matrix plotted for the Random Forest model. The plot shows that a high number of predictions were correctly classified. Out of 8 months predicted as having below average rainfall, 7 were correctly classified whereas only 1 was wrong classified. As well as all 4 months predicted as having Average rainfall were correctly classified. Overall, 11 out 12 predictions were correctly classified. This represents 91.66% of correctly classified rainfall predictions.

Figure 8. Confusion Matrix for Random Forest, first experiment



Logistic Regression

The classification report for Logistic Regression model similarly achieved high scores in precision, recall, and F1-score, resulting in an overall accuracy of approximately 91%. The confusion matrix showed that a high number of predictions were correctly classified. Out of the 8 months predicted as having below-average rainfall, 7 were correctly classified, while only 1 was incorrectly classified. Additionally, all 4 months predicted as having average rainfall were correctly classified. Overall, 11 out of 12 predictions were accurately classified, representing 91.66% of correctly classified rainfall predictions. At this stage of the experiments, the Logistic Regression model exhibited a performance comparable to that of the Random Forest model.

MLPClassifier Networks

The classification report for MLPClassifier model similarly achieved high scores in precision, recall, and F1-score, resulting in an overall accuracy of approximately 91%. The confusion matrix showed a high number of correctly classified predictions. Out of the 8 months predicted to have below-average rainfall, 7 were correctly classified, with only 1 misclassification. Similarly, all 4 months predicted to have average rainfall were accurately classified. In total, 11 out of 12 predictions were correct, representing 91.66% accuracy in rainfall predictions. At this experimental stage, the MLPClassifier model exhibited a performance comparable to that of the Random Forest and Logistic Regression models.

Xgboost

The classification report for Xgboost showcased a perfect score in precision, recall, and F1-score metrics. These results contributed to an overall accuracy of 100%, marking a significant achievement for the Xgboost model at this stage of the experiments. It is noteworthy that the Xgboost model outperformed the Random Forest, Logistic Regression, and MLPClassifier models in terms of precision, recall, and F1-score, underlining its superior performance in accurately classifying rainfall predictions during the phase where the training data remains without faults. Figure 9 shows the classification report for Xgboost.

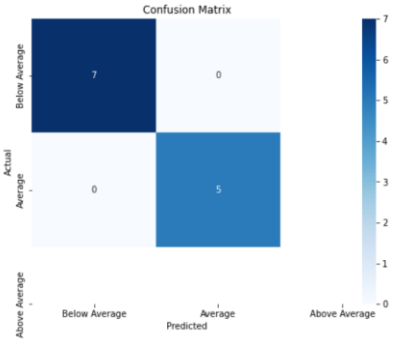
Figure 9. Classification report for Xgboost, first experiment

```
Accuracy = 1.0
Cohen's Kappa = 1.0
Time taken = 0.6256537437438965
              precision    recall  f1-score   support

           0    1.00000   1.00000   1.00000         7
           1    1.00000   1.00000   1.00000         5

    accuracy                        1.00000        12
   macro avg    1.00000   1.00000   1.00000        12
weighted avg    1.00000   1.00000   1.00000        12
```

The confusion matrix for Xgboost showed excellent performance. Particularly, all 7 months predicted to experience below-average rainfall were accurately classified, demonstrating a perfect record without any misclassifications. Similarly, the Xgboost model achieved perfection in classifying the 5 months predicted to have average rainfall. The overall outcome is notable, with a total of 12 out of 12 predictions being correct, resulting in 100% accuracy in forecasting rainfall. This stage of the experiments, characterized by fault-free training data, highlights the superior performance of the Xgboost model compared to the Random Forest, Logistic Regression, and MLPClassifier models.

Figure 10 below shows the confusion matrix plotted for the Xgboost model after the first experiment.

Figure 10. Confusion Matrix for Xgboost, first experiment



Performance Analysis – with faults introduced

In the subsequent phases of the experiments, deliberate faults were introduced into the training data to replicate the real-world challenge of missing data, as acknowledged by the Zambia Meteorology Department (ZMD). These introduced faults aimed to simulate scenarios where data gaps or inconsistencies could occur, allowing for a thorough evaluation of the models' adaptability to such challenges. Once the faults were incorporated, all four models—Xgboost, Random Forest, Logistic Regression, and MLPClassifier—were thoroughly examined using the same dataset. This evaluation sought to determine and compare their respective performances in the presence of faulty data, shedding light on their resilience and effectiveness under adverse conditions.

Random Forest

At this stage of the experiments, the performance of Random Forest model decreased from 91% to 83%. Despite this decline, the accuracy remains relatively high, showcasing the model's resilience to faulty data in the subsequent experiments.

Logistic Regression

The performance of Logistic Regression model decreased from 91% to 75%. This represents a sharper decline in accuracy than that seen in the Random Forest model. This suggests that the Logistic Regression model is less resilient to data faults than the Random Forest model.

MLPClassifier Networks

The model's overall accuracy decreased from 91% to 75%. This represents a sharper decline in accuracy than that seen in the Random Forest model. However, this performance at this stage is the same as seen in the Logistic Regression model. Overall, this shows that the MLPClassifier model, just like the Logistic Regression, is less resilient to data faults than the Random Forest model.

Xgboost

The model's overall accuracy decreased from 100% to 91%. Despite this decline, the accuracy remains relatively high, showcasing the model's resilience to faulty data in the second set of experiments. At this stage of the experiment, the performance of Xgboost is higher than that of the other models: Random Forest, Logistic Regression, and MLPClassifier.

Comparative Analysis

After the third round of experiments, where the faults introduced were doubled, a systematic comparative analysis was carried out. This related to the four models– Xgboost, Random Forest, Logistic Regression, and MLPClassifier. After this phase of experiments, it became evident that Xgboost and Random Forest exhibited superior resilience to the introduced faults. The results highlight a subtle relationship between the extent of data faults and model accuracy. As the proportion of faults increased, a discernible reduction in accuracy was observed across all models. However, Xgboost and Random Forest consistently outperformed the others even under heightened fault conditions. This robust performance underscores their efficacy in handling the challenges posed by data gaps and inconsistencies.
The is illustrated in table 2 below.

Table 2. Comparative analysis of models

	First Experiment	Second Experiment	Third Experiment
Model	Before fault injection	After fault injection	Fault injection doubled
Random Forest	91%	83%	66%
Logistic Regression	91%	75%	50%
MLPClassifier	91%	75%	50%
XGBoost	100%	91%	66%

Consequently, in situations where datasets exhibit significant data quality issues, the integration of external datasets, such as Reanalysis and Enact data, becomes instrumental in addressing and filling the prevalent data gaps. These external datasets serve as valuable supplements, enhancing the overall completeness and accuracy of the primary dataset. By leveraging reanalysis and Enact data, which amalgamate satellite and station data, the study proposes a comprehensive approach to fortifying the integrity of historical weather data, particularly in instances where the primary dataset suffers from substantial quality deficiency.

V. Conclusion

This study was undertaken for a comprehensive exploration into machine learning applications for seasonal rainfall forecasting, focusing on two pivotal research questions. The first question aimed at identifying resilient machine learning models in the face of poor data quality. Employing a systematic analytical approach and leveraging existing knowledge in robust algorithms, the study revealed insights into the performance and resilience of various models, with XGBoost and Random Forest emerging as particularly robust performers. The second research question delved into strategies for enhancing data quality using external datasets. Through a meticulous investigation of literature and practical data experiments, the study highlighted the efficacy of external sources like Reanalysis and Enact data, illustrating their potential to address data gaps and improve overall data quality. These findings collectively contribute to advancing the field of machine learning applications in seasonal rainfall forecasting. The insights gained provide valuable guidance for practitioners and meteorological departments, emphasizing the importance of model selection in handling data quality challenges. The study underscores the practical utility of external datasets in mitigating data gaps, offering tangible strategies for optimizing the accuracy and reliability of seasonal rainfall predictions. As meteorology increasingly relies on data-driven approaches, these outcomes become pivotal in shaping future methodologies and ensuring the resilience of machine learning models in the complex sphere of weather predictions.

References

[1] J. T. Esteves, G. De Souza Rolim, And A. S. Ferraudo, "Rainfall Prediction Methodology With Binary Multilayer Perceptron Neural Networks," Clim Dyn, Vol. 52, No. 3–4, Pp. 2319–2331, Feb. 2019, Doi: 10.1007/S00382-018-4252-X.

[2] L. Mzyece, M. Nyirenda, And J. Phiri, "Forecasting Seasonal Rainfall Using A Feed Forward Neural Network With Back-Propagation: A Case Of Zambia," Vol. 25, Pp. 8–18, Apr. 2024, Doi: 10.9790/0661-2506030818.

[3] F. Chulu, J. Phiri, M. Nyirenda, M. M. Kabemba, P. Nkunika, And S. Chiwamba, "Developing An Automatic Identification And Early Warning And Monitoring Web Based System Of Fall Army Worm Based On Machine Learning In Developing Countries," Zambia Ict Journal, Vol. 3, No. 1, Pp. 13–20, Mar. 2019, Doi: 10.33260/Zictjournal.V3i1.71.

[4] L. Mzyece, M. Nyirenda, M. K. Kabemba, And G. Chibawe, "Forecasting Seasonal Rainfall In Zambia – An Artificial Neural Network Approach," Zambia Ict Journal, 2018, [Online]. Available: Https://Api.Semanticscholar.Org/Corpusid:134095146

[5] T. Alemneh, "Rainfall Forecasting Using Support Vector Regression And Artificial Neural Network Models: A Case Study In The Upper Blue Nile Basin, Ethiopia," J Hydrol (Amst), Vol. 54, Pp. 429–443, 2017.

[6] A. Chan, N. Narayanan, A. Gujarati, K. Pattabiraman, And S. Gopalakrishnan, "Understanding The Resilience Of Neural Network Ensembles Against Faulty Training Data," In 2021 Ieee 21st International Conference On Software Quality, Reliability And Security (Qrs), Ieee, Dec. 2021, Pp. 1100–1111. Doi: 10.1109/Qrs54544.2021.00118.

[7] S. Sumi, M. F. Zaman, And H. Hirose, "A Rainfall Forecasting Method Using Machine Learning Models And Its Application To The Fukuoka City Case," International Journal Of Applied Mathematics And Computer Science, 2012, Doi: 10.2478/V10006-012-0062-1.

[8] G. T. Ferrari And V. Ozaki, "Missing Data Imputation Of Climate Datasets: Implications To Modeling Extreme Drought Events," Revista Brasileira De Meteorologia, Vol. 29, No. 1, Pp. 21–28, Mar. 2014, Doi: 10.1590/S0102-77862014000100003.

[9] M. A. Holmstrom And D. Liu, "Machine Learning Applied To Weather Forecasting," 2016. [Online]. Available: Https://Api.Semanticscholar.Org/Corpusid:12439970

[10] S. Huber, H. Wiemer, D. Schneider, And S. Ihlenfeldt, "Dmme: Data Mining Methodology For Engineering Applications – A Holistic Extension To The Crisp-Dm Model," Procedia Cirp, 2019, Doi: 10.1016/J.Procir.2019.02.106.

[11] S. Jaggia, A. F. Kelly, K. Lertwachara, And L. Chen, "Applying The Crisp-Dm Framework For Teaching Business Analytics," Decision Sciences Journal Of Innovative Education, 2020, Doi: 10.1111/Dsji.12222.

[12] F. R. Kurniawan And R. Sutomo, "Forecasting Rice Inventory In Indonesia Using The Arima Algorithm Method," Journal Of Multidisciplinary Issues, 2021, Doi: 10.53748/Jmis.V1i2.15.

[13] H. J. Gómez Palacios, R. A. Jiménez Toledo, G. Hernández, And Ã. A. Martínez Navarro, "A Comparative Between Crisp-Dm And Semma Through The Construction Of A Modis Repository For Studies Of Land Use And Cover Change," Advances In Science Technology And Engineering Systems Journal, 2017, Doi: 10.25046/Aj020376.

[14] S. J. Koopman And K. M. Lee, "Seasonality With Trend And Cycle Interactions In Unobserved Components Models," J R Stat Soc Ser C Appl Stat, Vol. 58, No. 4, Pp. 427–448, Sep. 2009, Doi: 10.1111/J.1467-9876.2009.00661.X.

[15] F. Martínez-Plumed Et Al., "Crisp-Dm Twenty Years Later: From Data Mining Processes To Data Science Trajectories," Ieee Trans Knowl Data Eng, 2021, Doi: 10.1109/Tkde.2019.2962680.

[16] R. S. Schumacher Et Al., "From Random Forests To Flood Forecasts: A Research To Operations Success Story," Bull Am Meteorol Soc, 2021, Doi: 10.1175/Bams-D-20-0186.1.

[17] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, And Y. Si, "A Data-Driven Design For Fault Detection Of Wind Turbines Using Random Forests And Xgboost," Ieee Access, 2018, Doi: 10.1109/Access.2018.2818678.

[18] R. Praveena, T. R. Babu, M. Birunda, G. Sudha, P. Sukumar, And J. Gnanasoundharam, "Prediction Of Rainfall Analysis Using Logistic Regression And Support Vector Machine," J Phys Conf Ser, 2023, Doi: 10.1088/1742-6596/2466/1/012032.

[19] L. Anselin, S. Sridharan, And S. Gholston, "Using Exploratory Spatial Data Analysis To Leverage Social Indicator Databases: The Discovery Of Interesting Patterns," Soc Indic Res, Vol. 82, No. 2, Pp. 287–309, Jun. 2007, Doi: 10.1007/S11205-006-9034-X.

[20] P. A. Sukhonos And N. A. Diansky, "The Role Of Winter Net Heat Fluxes On The Modulation Of The Upper Mixed Layer Temperature And Depth In The North Atlantic By The Reanalysis Data," Iop Conf Ser Earth Environ Sci, 2022, Doi: 10.1088/1755-1315/1040/1/012032.