

*An internship report submitted in partial fulfilment of the requirements for the*

***Award of Degree of***

### BACHELOR OF TECHNOLOGY

**In**

### COMPUTER SCIENCE AND ENGINEERING (CYBER SECURITY)

#### By

**ROSHNI VENKATESAN**

#### Roll No: CB.EN.U4CYS21061

***“A Lightweight ELK-Based Modular SIEM Pipeline For Real Time Threat Intelligence Using Distributed Log Correlation”***

**Under Supervision of**

### R. DEVENDAR NAIK

### DEPUTY MANAGER, SCOF/RO

**SDSC-SHAR, SRIHARIKOTA, ANDHRA PRADESH, 524124**

(Duration: 24th April 2025 – 23rd May 2025)

TIFAC-CORE IN CYBERSECURITY

**AMRITA VISHWA VIDYAPEETHAM COIMBATORE**

**AMRITA NAGAR P.O., ETTIMADAI,**

**TAMIL NADU - 641112**

# ACKNOWLEDGEMENT

I feel privileged and thankful to mention such esteemed dignitaries who aided me for successful completion of internship training. My best regards to **Shri G. GRAHADURAI, Deputy Director, RO**, for giving permission for doing internship at Range Operation.

My best regards to **Srimathi. V. LATHA, General Manager, SCOF RO**, for her support throughout the Internship at Range Operations.

I express my sincere gratitude to **Shri. DEVENDAR NAIK, DEPUTY MANAGER, SCOF/RO** whose support and belief in me laid stepping stones in accomplishment of my project work. I would like to thank **Shri. NANDA SUVAN, SCIENTIST /ENGINEER - SC, SCOF/RO** for sharing guidance.

I express my sincere thanks and deep veneration to **Prof. M. SETHUMADHAVAN**, Centre Head at **AMRITA VISHWA VIDYAPEETHAM, COIMBATORE** for giving me this opportunity for successful completion of internship training.

In conclusion, we shall remember our project training and put an oath of presenting the training experience to prove our ability and work for the pride of the organization in all respects wherever we work.

**ROSHNI VENKATESAN**

**CB.EN.U4CYS21061**

# BONAFIDE CERTIFICATE

This is to certify that **ROSHNI VENKATESAN**, a student at **AMRITA VISHWA VIDYAPEETHAM, COIMBATORE ,** CB.EN.U4CYS21061 has completed her

project training successfully at **SHAR COMPUTER FACILITIES** (SCOF)/RO, SATISH DHAWAN SPACE CENTRE, SRIHARIKOTA, ANDHRA PRADESH, 524124 from 24-04-2025 to 23-05-2025.

During the internship period his conduct was found to be .

**V. LATHA DEVENDAR NAIK** SCIENTIST/ENGINEER-SG, SCIENTIST/ENGINEER-SE, GENERAL MANAGER PROJECT GUIDE

SCOF, RO SCOF, RO

# INDEX

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Content** | **Page No.** |
| 1 | Acknowledgement | 2 |
| 2 | 1. About the Organization    1. Introduction    2. Range Operations (RO)    3. SHAR Computer Facility (SCOF) Operations | 6-11 |
| 3 | 1. Technology Used    1. Python    2. Anaconda    3. MySQL    4. MySQL Workbench 8.0 CE | 12-18 |
| 4 | 1. Project Details    1. Overview    2. Related Work    3. Flow Chart    4. Data Preprocessing    5. Implementation and Results    6. Analysis | 19-43 |
| 5 | Conclusion | 44-45 |
| 6 | Reference | 46 |

**ABSTRACT**

***“A Lightweight ELK-Based Modular SIEM Pipeline For Real Time Threat Intelligence Using Distributed Log Correlation”***

This project presents a lightweight ELK-based modular SIEM pipeline aimed at delivering real-time threat intelligence through distributed log correlation and visualization. The primary goal is to unify and visualize massive log data from diverse sources such as Windows Event Viewer, Active Directory, and firewall logs. By leveraging Beats for endpoint data collection, Logstash for parsing and enrichment, Elasticsearch for structured storage, and Kibana for interactive dashboards, the system enables immediate insights into security-relevant activities.

Unlike traditional architectures that focus on post-incident log analysis, the proposed approach empowers users to write custom rules for monitoring parameters such as RDP request volumes, admin logins, privileged activities, blocked requests, suspicious event codes, and traffic spikes. This modular design supports real-time detection and visualization of anomalies, allowing for proactive security monitoring. Inputs are structured as key-value pairs, and the system supports flexible rule creation for event codes, protocols, IP tracking, and denied requests, all visualized through customizable Kibana dashboards.

Performance evaluation demonstrates that the pipeline delivers low-latency, real-time log visualization with secure, role-based access for multiple users. The results highlight a significant improvement in situational awareness and threat detection compared to conventional post-exploit log searches. Overall, this work underscores the value of modular, distributed log correlation in providing scalable, actionable security intelligence for modern enterprise environments.

# CHAPTER-1 ABOUT THE ORGANIZATION

## Introduction: -

The Satish Dhawan Space Centre SHAR (SDSC SHAR) in Sriharikota, known as the Spaceport of India, stands as a prominent hub within the Indian Space Research Organisation (ISRO), operating under the aegis of the Department of Space (DOS), Government of India. ISRO functions as an autonomous entity, playing a pivotal role in advancing India's space exploration endeavours.

Fig1.1 Satish Dhawan Space Centre SHAR, Sriharikota

Situated in Sriharikota, India, the SDSC SHAR is recognized globally as a leading rocket launch station. This state-of-the-art facility serves as a crucial nexus for launching a diverse array of space missions. These missions span a spectrum of objectives, including the deployment of satellites for remote sensing, communications, navigation, and scientific research.

The centre’s commitment extends to ensuring reliable access to space for both indigenous and commercial satellites, contributing to its esteemed reputation on the global stage. The selection of Sriharikota Island in 1969 marked the genesis of this rocket launch station. Since its operational debut on October 9, 1971, with the launch of the Rohini sounding rocket, the center has continually expanded its facilities to meet the dynamic requirements of ISRO.

SDSC SHAR's operational activities are intricately categorized into various domains. These include Vehicle Assembly and Static Test Operations, Range Operations, Liquid Storage and Service Facilities, Solid Propellant, and the Space Booster Plant. These multifaceted operations are integral to supporting the launch of indigenously designed and developed vehicles, such as SLV, ASLV, PSLV, GSLV Mk-II, and GSLV Mk-III.



Fig:1.2 Chandrayaan-3 Launching Fig:1.3 GSLV-MkII

Beyond its role in launching vehicles, the center takes charge of critical aspects such as program planning, human resources development, and the oversight of systems reliability groups. Administrative and auxiliary support for the center is seamlessly facilitated through the Sriharikota Common Facilities, completing the comprehensive infrastructure that underscores SDSC SHAR's pivotal role in advancing India's prowess in space exploration.

## Range Operations: -

The Range Operations Entity serves as the central hub for overseeing launch operations across various missions conducted by the Indian Space Research Organisation (ISRO) at the Satish Dhawan Space Centre (SDSC) SHAR.



Fig1.4 Range Operations

This entity is entrusted with a diverse array of responsibilities, spanning tracking, tele-commanding, real-time systems for mission monitoring, and the deployment and maintenance of various networks, including the mission network, campus network, internet network, and surveillance network. It also manages networking services with secure features, computerization of administrative activities, web and mobile application development, as well as photography, including still and video coverage. Additionally, it provides meteorology services, ensuring constant weather monitoring to facilitate the clearance of various launch campaign activities.

#### Key operations of Range Operations at SDSC:-

* + 1. **Launch Coordination:**

Range Operations involve meticulous coordination of the launch activities, ensuring that all systems and components are functioning optimally before the launch sequence begins.

#### Tracking and Telemetry:

The range includes a network of tracking stations equipped with radar and telemetry systems to monitor the trajectory and performance of the launch vehicle throughout its journey. This real-time data is crucial for assessing the mission's success and making necessary adjustments.

#### Safety Protocols:

Range safety is paramount during launch operations. Rigorous safety protocols are established to protect personnel, the environment, and surrounding areas. These protocols include criteria for trajectory deviations, ensuring that the launch vehicle follows its planned path.

#### Abort Criteria:

Range Operations define criteria for mission abort or vehicle destruction in case of any anomalies that could jeopardize the mission or public safety. These criteria are established to minimize risks associated with launch failures.

#### Communication Systems:

The range is equipped with robust communication systems to maintain constant contact with the launch vehicle. This includes voice and data communication to provide real-time updates and receive commands.

#### Flight Safety Analysis:

Prior to each launch, extensive flight safety analysis is conducted to assess potential risks and determine the necessary precautions. This analysis considers various scenarios to ensure the safety of the mission.

#### Post-Launch Analysis:

After the launch, Range Operations continue with the analysis of post-flight data. This involves assessing the performance of the launch vehicle and its systems, identifying any anomalies, and applying lessons learned to enhance future missions.

#### International Collaboration:

ISRO's Range Operations may involve collaboration with international space agencies or organizations for tracking and telemetry support, enhancing the global network of space surveillance and monitoring.

The Range Operations Entity houses crucial facilities such as the Mission Control Centre, Computer and Communications Centre, Multi-Object Tracking Radar (MOTR), including mobile MOTR, and the SHAR Computer Facilities Operations (SCOF).

## SHAR Computer Facility (SCOF) Operations: -

The SHAR Computer Facility meticulously addresses the computational requirements of the Center by unifying hardware, software, and networking necessities. Adhering to stringent security guidelines, the facility has implemented a segregation between the Internet and Intranet.

Noteworthy initiatives undertaken by SCOF include the establishment of a Network Operations Centre, which also serves the dual role of a Data Centre. State-of-the-art High-Performance Computing (HPC) servers have been strategically commissioned to cater to diverse needs, including meteorological and MOTR requirements within the Center.

Furthermore, the facility has successfully implemented three real-time systems as integral components of the mission network: the Range Safety (RS) real-time system, the Specialists’ Display System (SDS) real-time system, and the Mission Control Centre (MCC) real-time system.



Fig:1.5 Mission control centre

These systems play a pivotal role in continuously monitoring the status of various ground stations leading up to launch. Additionally, they facilitate real time monitoring during the critical filling phase of the liquid and cryogenic stages of the launch vehicle in the countdown phase. Post-lift-off, these systems prove instrumental in monitoring the performance of vehicle subsystems.

# CHAPTER-2

**TECHNOLOGY USED**

## ELK Stack: -

The ELK Stack is a combination of three open-source tools—Elasticsearch, Logstash, and Kibana—designed to help users collect, search, analyze, and visualize large volumes of data from multiple sources. Logstash ingests and processes data, Elasticsearch stores and indexes it for fast search and analytics, and Kibana provides interactive dashboards and visualizations for real-time insights. This stack is widely used for centralized log management, security analytics, and infrastructure monitoring, making it easier to detect issues, monitor systems, and gain valuable operational intelligence.

#### 2.1.1 Elasticsearch:

Elasticsearch is a powerful search and analytics engine that is used to store and manage very large amounts of data. It is designed to help people quickly find and analyze information from huge collections of records, such as logs or documents. When data is sent to Elasticsearch, it organizes the information so that you can search for specific details, filter results, and even run complex queries to spot patterns or trends. It is widely used because it can handle a lot of data at once and still provide answers in just a few seconds.

#### 2.1.2 Logstash:

Logstash is a tool that helps collect, process, and prepare data before it is stored. It acts like a middleman that takes in raw data from many different sources, cleans it up, and changes it into a format that is easier to work with. Logstash can filter out unnecessary information, add extra details, and organize everything so that the data makes sense and is ready for searching or analysis. It is very flexible and can be set up to handle many types of data and different processing rules.

#### 2.1.3 Kibana:

Kibana is a visualization tool that makes it easy to see and understand data stored in Elasticsearch. It provides a web-based interface where users can create charts, graphs, and dashboards to display information visually. With Kibana, people can explore their data, spot trends, and monitor key activities using interactive visuals. It is especially helpful for turning complex sets of data into clear and understandable pictures that can be shared with others.

## Beats: -

Beats are small, lightweight programs that are installed on computers or devices to collect and send data to other systems for storage or analysis. They are designed to be simple and efficient, so they do not use much computer power or memory. Beats are often used to gather logs, metrics, or network information from different machines and send this data to tools like Logstash or Elasticsearch.

#### 2.2.1 Winlogbeat:

Winlogbeat is a lightweight, open-source agent designed to collect and ship Windows event logs—including application, security, system, and hardware events—from Windows systems to destinations like Elasticsearch or Logstash for storage and analysis. It runs as a Windows service, continuously monitors specified event logs, and forwards new event data in near real time, making it easier to centralize and analyze Windows logs across multiple machines.

#### 2.2.2 Syslog:

A syslog server built into FortiGate is a feature that allows the firewall to collect, store, and forward log messages about network activity, security events, and system operations. It uses the standard syslog protocol to send these logs to external servers for centralized monitoring and analysis. Administrators can configure which types of events and severity levels are logged, logs can be sent over different ports and protocols. This capability helps organizations track network behavior, detect threats, and maintain records for auditing and compliance.

## Infrastructure: -

#### 2.3.1 Ubuntu 24.04:

Ubuntu 24.04 is a modern version of the popular Linux operating system. It is known for being stable, secure, and user-friendly, making it a common choice for running servers and desktop computers. Ubuntu provides a flexible environment for installing software, managing files, and connecting to networks. It supports both graphical and command-line interfaces, allowing users to interact with the system in the way that suits them best. Regular updates and strong community support help keep Ubuntu systems reliable and up to date.

#### 2.3.2 Active Directory:

Active Directory is a directory service developed by Microsoft that helps organizations manage users, computers, and other resources on a network. It acts as a central database where information about user accounts, permissions, and networked devices is stored. Active Directory makes it easier for administrators to control access, enforce security policies, and keep track of changes within the network. It is widely used in business and educational environments to organize and secure digital resources.

#### 2.3.3 Event Viewer:

Event Viewer is a built-in tool in Microsoft Windows that records detailed information about system events, application errors, security incidents, and other important activities on a computer. It allows users and administrators to review logs, troubleshoot problems, and monitor the health and security of Windows systems. By examining events in the Event Viewer, it is possible to identify issues, track user actions, and ensure that the system is operating as expected.

#### 2.3.4 Fortigate Firewall:

A FortiGate firewall is a security device that protects computer networks from unauthorized access and cyber threats. It monitors all incoming and outgoing network traffic, blocks harmful connections, and provides detailed logs about network activity. The firewall can enforce security rules, detect suspicious behavior, and help prevent attacks on the network. FortiGate firewalls are commonly used by organizations to maintain a safe and secure network environment.

# CHAPTER-3

**PROJECT DETAILS**

## Overview: -

Accurate weather prediction is critical for applications such as agriculture, water resource management, disaster preparedness, and climate research. Traditional forecasting models often struggle with the complex, nonlinear interactions within atmospheric data. Recent advances in machine learning (ML) and deep learning (DL) have opened new avenues by incorporating both spatial and temporal dynamics. In our project, we compared a range of models—including LSTM, CNN-LSTM, Temporal Convolutional Networks (TCN), transformers, XGBoost, and hybrid approaches (XGBoost+LSTM and XGBoost+Random Forest)—to identify the optimal framework for precise weather forecasting.

#### Role of LSTM in Weather Prediction:

Designed for time-series data, LSTM networks are well-suited to capture long-term dependencies in historical weather data, such as trends in temperature, humidity, and wind conditions. They offer robust handling of noisy or missing data, which is common in meteorological records.

#### Role of CNN-LSTM in Weather Prediction:

By integrating CNN with LSTM, the CNN-LSTM model effectively extracts spatial features from satellite imagery and radar data while simultaneously modeling temporal sequences. This combination enhances the model's ability to detect evolving weather patterns that influence forecasts.

#### Role of TCN (Temporal Convolutional Network):

TCN leverages convolutional architectures to process sequential data, offering a competitive alternative to recurrent models. They are particularly effective in capturing temporal dependencies through parallel processing, leading to improved efficiency and accuracy in weather prediction.

#### Role of Transformers:

Utilizing self-attention mechanisms, transformer models excel at identifying long-range dependencies in time-series data. Their capacity to dynamically focus on different aspects of the input data makes them well-suited for modeling the intricate and variable nature of atmospheric processes.

#### Role of XGBoost:

As a gradient boosting framework, XGBoost is highly effective with structured, tabular data. It excels at handling diverse meteorological features (e.g., temperature, humidity, wind speed, pressure) and provides valuable insights through feature importance analysis.

#### Role of Hybrid Approaches (XGBoost+LSTM and XGBoost+Random Forest):

The hybrid approaches combine the strengths of deep learning with the interpretability and robustness of ensemble methods. By merging the feature extraction capabilities of LSTM with the precision of XGBoost or Random Forest, these models offer enhanced performance by capturing complex nonlinear relationships in weather data.

This comprehensive evaluation of different modeling techniques highlights the importance of integrating multiple approaches to effectively capture the spatial and temporal dynamics of atmospheric phenomena. Our findings indicate that hybrid methods, particularly the combination of XGBoost and Random Forest, provide a promising direction for advancing the accuracy and reliability of weather forecasting systems.

## Related Work: -

Weather prediction has long been a challenging task due to the dynamic and complex nature of atmospheric processes. Traditional forecasting methods, such as numerical weather prediction (NWP) and statistical models, often struggle to fully capture the intricate non-linear relationships governing weather patterns. In recent years, the advent of machine learning (ML) and deep learning (DL) has enabled more sophisticated approaches to weather forecasting. While LSTM has been extensively used for time-series forecasting due to its ability to retain long-term dependencies, our study extends beyond a single-model approach to a comparative evaluation of multiple advanced architectures. Specifically, our investigation includes LSTM, CNN-LSTM, Temporal Convolutional Networks (TCN), transformers, and hybrid methodologies such as XG-Boost with LSTM and XG-Boost combined with Random Forest. By evaluating these diverse models, we aim to determine the most effective technique for capturing the non-linear dependencies inherent in meteorological data.

Rainfall/Precipitation forecasting, as a fundamental aspect of weather prediction, involves understanding key meteorological concepts:

* + - **Precipitation** refers to any form of water (rain, snow, sleet, or hail) falling from the atmosphere to the Earth's surface.
    - **Rainfall Intensity** is measured in millimeters per hour (mm/h) and plays a crucial role in flood prediction and water resource management.
    - **Time-Series Forecasting** is the process of using historical weather data to predict future rainfall trends.
    - **Spatiotemporal Analysis** is essential for capturing variations in weather data across both time and space, allowing models to recognize regional patterns of precipitation.

By systematically evaluating different machine learning and deep learning architectures, our research contributes to the ongoing advancement of precipitation forecasting. The findings underscore the importance of diversifying model selection, optimizing feature engineering, and leveraging hybrid learning techniques to develop more accurate and reliable weather prediction systems. These insights provide a strong foundation for future studies aimed at refining precipitation models, integrating real-time meteorological data, and further enhancing forecasting precision for practical applications in disaster management, agriculture, and climate research.

## Flow Chart: -

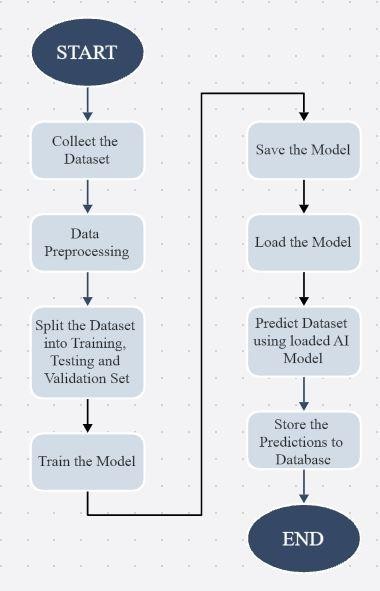
****

Fig. 3.1 Flowchart of the Program

## Data Preprocessing: -

Data preprocessing is a crucial step in ensuring that weather forecasting models are trained on clean, well-structured, and meaningful data. This phase involves collecting meteorological data, handling missing values, standardizing formats, and structuring the dataset for efficient model training. Proper preprocessing enhances model accuracy by removing noise and inconsistencies while ensuring that relevant features are effectively utilized.

#### Data Collection:

Relevant meteorological data is gathered from multiple sources, including:

* + - **Satellite Imagery:** Cloud cover patterns and atmospheric conditions extracted from satellite-based observations.
    - **Reanalysis Datasets:** Historical weather data compiled using numerical weather prediction models (e.g., ERA5, IMERG).
    - **Radar Data:** Precipitation intensity and spatial distribution captured using radar reflectivity measurements.
    - **Sensor-Based Observations:** Data collected from ground-based weather stations, including temperature, humidity, wind speed, and atmospheric pressure.
    - **API Integrations:** Real-time weather data obtained from services like NOAA, IMD, or OpenWeatherMap for continuous model updates.

## Implementation and Results: -

The implementation phase focuses on training and evaluating multiple machine learning and deep learning models for rainfall prediction. We experimented with architectures such as LSTM, CNN-LSTM, TCN, Transformers, and hybrid models like XGBoost with LSTM and Random Forest. Each model was trained using preprocessed meteorological data, with hyperparameter tuning applied for optimization. The results were analyzed using performance metrics such as RMSE, MAE, and R² score, allowing for a comparative assessment of model effectiveness. Our findings highlight the advantages of hybrid ensemble methods in improving predictive accuracy, demonstrating their potential for robust weather forecasting.

#### Experiment setup:

Our study utilized a comprehensive experimental setup to evaluate multiple machine learning and deep learning models for rainfall prediction.

* + - **Dataset:** Historical meteorological data, including temperature, humidity, wind speed, pressure, and additional rainfall-related features.
    - **Features Used:** Seven core meteorological features (temperature, humidity, wind direction, wind speed, pressure, zonal wind, meridional wind) along with separately tested rainfall and precipitation data.
    - **Input Window:** 1 Hour.
    - **Output Window:** 3 Hours.
    - **Training Period:** 2000–2012.
    - **Testing Period:** 2012–2016.
    - **Validation Period:** 2016–2020.
    - **Evaluation Metrics:** Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²) score.

We compared models including LSTM, CNN-LSTM, Temporal Convolutional Networks (TCN), Transformers, XGBoost, and hybrid approaches such as XGBoost combined with LSTM and XGBoost with Random Forest. The hybrid XGBoost-Random Forest model demonstrated superior predictive performance, effectively capturing non-linear dependencies in the data. The results highlight the potential of ensemble methods in enhancing forecast accuracy. Future research can focus on refining hyperparameters, incorporating additional atmospheric variables, and extending the training period to improve model generalization.

#### Comparison of Rainfall Prediction Models:

Accurate weather prediction plays a vital role in various sectors, including agriculture, disaster management, and climate research. This study conducts a comparative analysis of multiple advanced machine learning and deep learning models to determine the most effective approach for rainfall forecasting. We evaluate Long Short-Term Memory (LSTM), Convolutional Neural Networks combined with LSTM (CNN-LSTM), Temporal Convolutional Networks (TCN), Transformers, and ensemble methods such as XGBoost-LSTM and XGBoost-Random Forest. The objective is to identify which model best captures the complex spatiotemporal dependencies of meteorological data while ensuring high accuracy and reliability. Through extensive experimentation, we assess model performance based on multiple evaluation metrics, highlighting the advantages of hybrid and ensemble techniques over standalone deep learning models.

#### LSTM(Long Short-Term Memory):

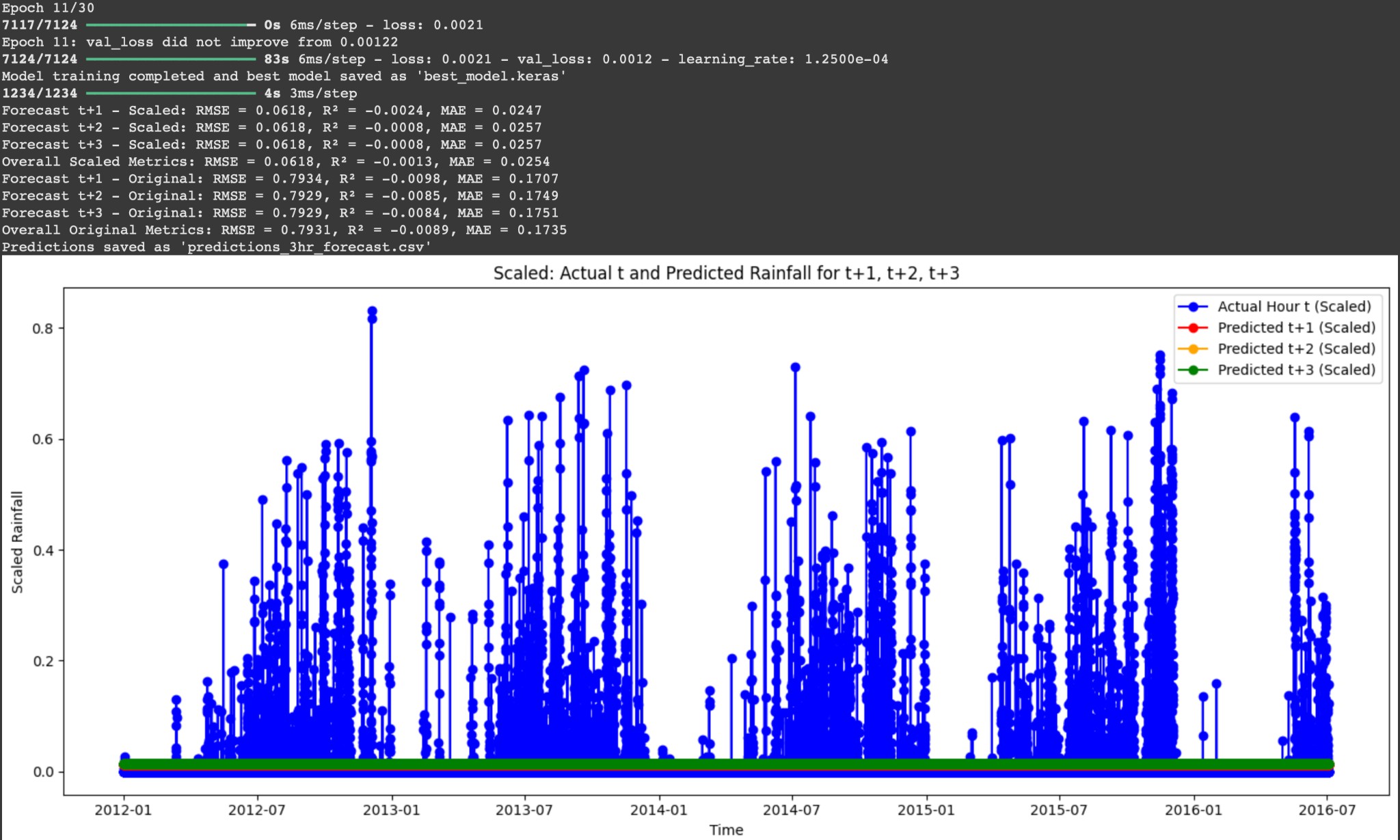
LSTM is a type of Recurrent Neural Network (RNN) designed to handle time-series data. It captures long-term dependencies and patterns in sequential data, making it suitable for rainfall prediction. By using memory cells and gating mechanisms, LSTM mitigates issues like vanishing gradients, ensuring better retention of past information over long sequences.

Fig. 3.2 LSTM Metrics and 3-Hour Forecast Plot

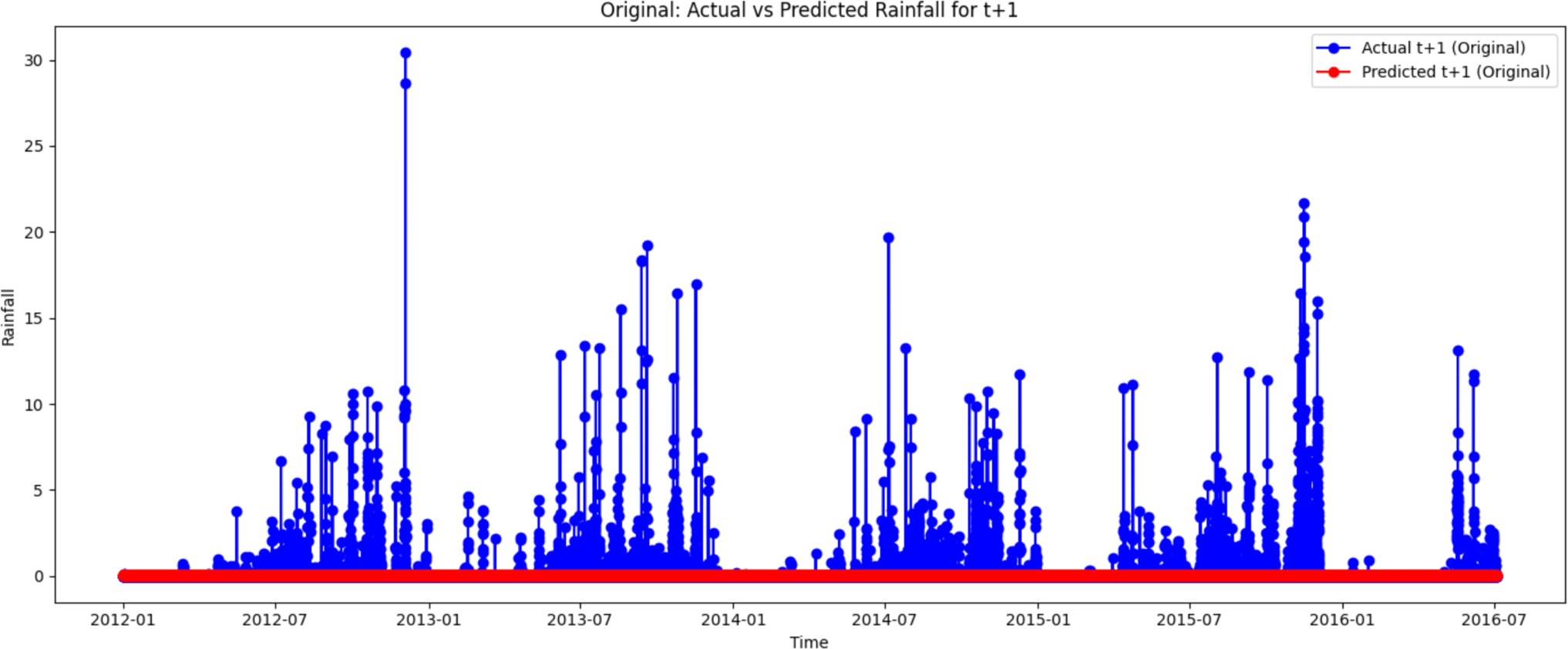


Fig. 3.3 LSTM 1-Hour Forecast Plot

#### Best Model’s Precipitation Prediction:

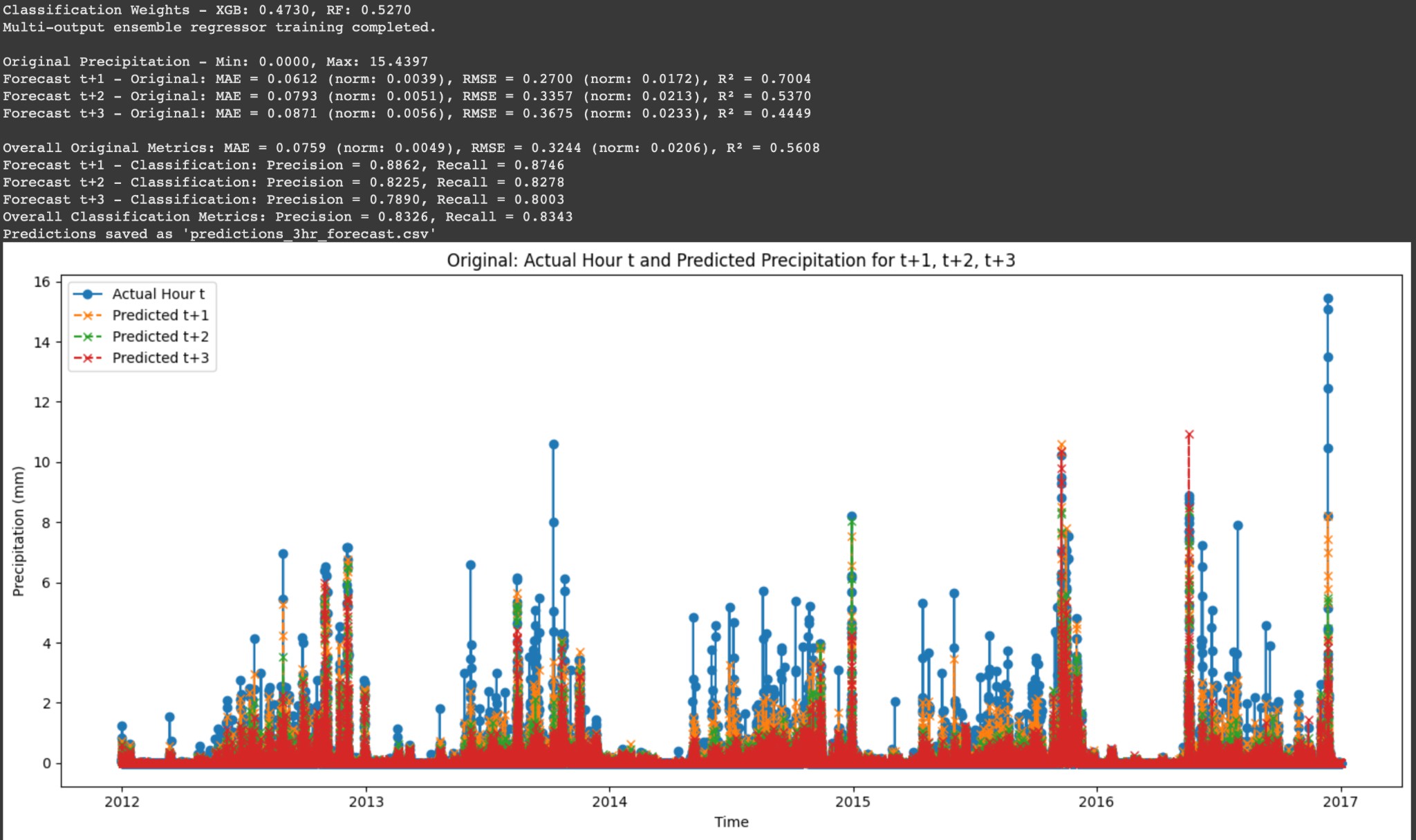
****

Fig. 3.16 Metrics and 3-Hour Forecast Plot for Precipitation using XGBoost-RF

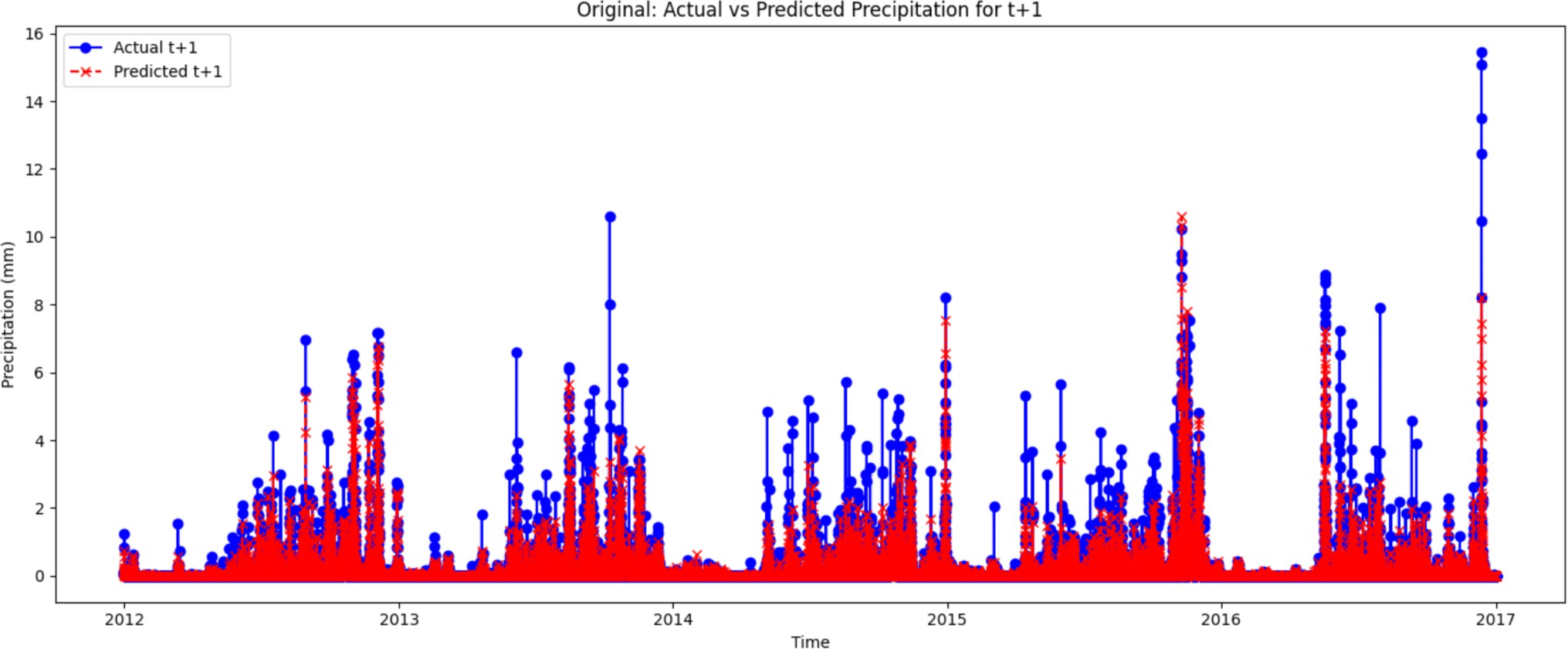


Fig. 3.17 1-Hour Forecast Plot for Precipitation using XGBoost-RF

## Analysis: -

The evaluation of various machine learning models for predicting rainfall and precipitation has provided valuable insights, highlighting the strengths and limitations of different approaches. Each model demonstrated varying levels of accuracy in capturing temporal and spatial patterns, emphasizing the importance of selecting the most suitable architecture for weather forecasting.

#### Prediction Output for Rainfall:

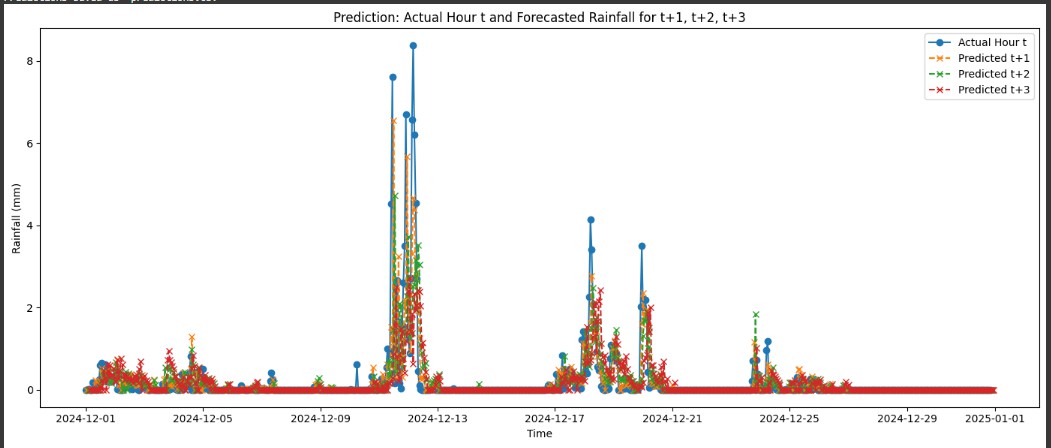
****

Fig. 3.18 Prediction of 3-Hour Forecast Plot for Rainfall

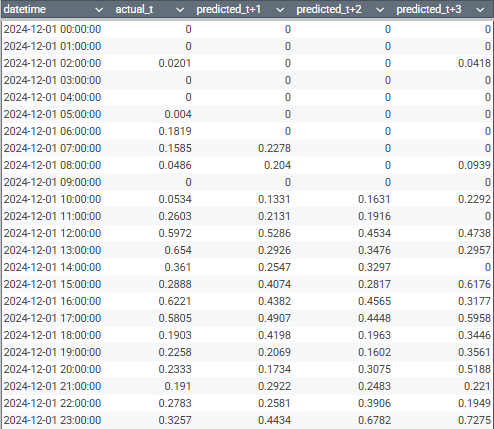


Fig. 3.19 Prediction of 3-Hour Forecast Values for Rainfall

#### Comparative Analysis for Rainfall/Precipitation Prediction:

Rainfall and precipitation forecasting play a critical role in sectors such as agriculture, disaster management, and climate research. Machine learning models have been increasingly utilized to enhance prediction accuracy by capturing complex atmospheric patterns. This study presents a comparative analysis of Long Short-Term Memory (LSTM), Convolutional Neural Networks with LSTM (CNN-LSTM), Temporal Convolutional Networks (TCN), Transformers, and hybrid models such as XGBoost-LSTM and XGBoost-Random Forest for weather prediction. The findings reveal that the hybrid XGBoost-Random Forest model outperforms other approaches by leveraging both ensemble learning and sequential pattern recognition.

* **Source:** Meteorological data from weather stations and satellite records.
* **Features:** Temperature, humidity, wind speed, wind direction, pressure, zonal wind, meridional wind, with additional evaluation of rainfall and precipitation.
* **Time Window:** Historical time-series data used for multi-step forecasting.
* **Evaluation Metrics:** RMSE, MAE, R²-score.
* **Model Descriptions:**

### LSTM:

* + **Architecture:** Deep recurrent neural network.
  + **Strength:** Effective in capturing long-term dependencies in sequential data.
  + **Weakness:** Computationally intensive and requires careful hyperparameter tuning.

By leveraging diverse machine learning techniques, this study provides an in-depth comparison of different forecasting approaches, emphasizing the benefits of hybrid models in improving weather prediction accuracy.

# CHAPTER-4

**CONCLUSION**

Our project focused on exploring and comparing a range of machine learning models to predict rainfall and precipitation, including LSTM, CNN-LSTM, TCN, Transformer, XGBoost, XGBoost-LSTM, and XGBoost-Random Forest. After a comprehensive evaluation, the hybrid XGBoost-Random Forest model emerged as the top performer, showcasing its superior ability to capture the complex, non-linear patterns within the meteorological data. This model's performance highlighted the power of combining traditional machine learning techniques with ensemble methods to improve forecast accuracy.

Throughout the project, we focused on incorporating both conventional meteorological features and additional atmospheric data to create a more robust predictive model. By optimizing the data preprocessing, feature engineering, and model tuning steps, we were able to refine the system and improve the model's ability to adapt to varying weather conditions. These improvements were key in enhancing the forecasting performance and ensuring the model could handle diverse and unpredictable data patterns.

A key takeaway from this project is the advantage of using hybrid models over standalone ones. While individual models like LSTM and Transformer performed well on their own, the integration of XGBoost with Random Forest proved more effective by leveraging the strengths of both gradient boosting and decision tree methods. This insight underscores the importance of hybrid approaches in real-world problems, where no single model may be able to capture all the nuances in the data.

On a personal level, my time at SHAR Computer Facilities (SCOF) has been a rewarding and transformative experience. I gained hands-on knowledge in machine learning, Python programming, and database management using MySQL.

In addition, working on projects provided me with a broader technical skill set and practical experience. The mentorship I received from software developers was invaluable, helping me grow both professionally and personally. This experience reinforced the importance of time management, clear communication, and self-motivation in successfully completing complex projects.

Looking ahead, the results of this project provide a strong foundation for future advancements in weather prediction systems. There’s significant potential to further refine the models, integrate additional data sources such as satellite imagery or climate data, and explore more sophisticated techniques to continually improve accuracy. The insights gained from this work not only contribute to academic research but also have the potential to benefit industries like agriculture, disaster management, and transportation that rely on accurate weather forecasting.

# References

1. **Using the ELK Stack for SIEM** by Daniel Berman at *logz.io*
2. **SIEM Implementation in 4 Steps** by Steve Moore at *Exabeam*
3. **TryHackMe | Investigating with ELK 101** by igor\_sec in *Medium*
4. **Threat Hunting: Log Monitoring Lab Setup with ELK** by Raj *at Hacking Articles*
5. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017).** Attention is all you need. *In Advances in neural information processing systems* (pp. 5998-6008).
6. **Chen, T., & Guestrin, C. (2016).** XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
7. **Zhou, Z.-H. (2012).** Ensemble Methods: Foundations and Algorithms. *Chapman & Hall/CRC Press*.
8. **Breiman, L. (2001).** Random forests. *Machine learning, 45*(1), 5-32.
9. **Bastani, M., & Zepeda, M. (2020).** Predicting rainfall with machine learning.

*Scientific Reports, 10*(1), 19834.

1. **Li, J., Li, Z., & Zhang, H. (2020).** A hybrid deep learning model for rainfall prediction based on LSTM and Random Forest. *Computers, Environment and Urban Systems, 80*, 101426.