NETWORK PACKET PREDICTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

MANOJ R L (2116220701161)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **“NETWORK PACKET PREDICTION ”** is the bonafide work of **“MANOJ R L (2116220701161)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy, M.Tech., Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and
Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on ______________________

Internal Examiner **External Examiner**

ABSTRACT

FlowMeter is a machine learning-based network traffic classification system designed to assist cybersecurity professionals in identifying malicious activity within network flows. It processes flow-level statistics generated by monitoring tools and leverages predictive models to differentiate between benign and malicious behavior. The core objective of FlowMeter is to enhance intrusion detection capabilities through automation and intelligent analysis.

The system works by extracting key flow features such as packet size, duration, inter-arrival times, and throughput metrics, which are then preprocessed and fed into a logistic regression model. The model was trained and evaluated on real network traffic datasets containing both normal and attack flows. A hyperparameter tuning strategy and class weight balancing were employed to improve model accuracy and detection performance.

FlowMeter's backend pipeline is developed using Python and Scikit-learn, while the datasets used stem from simulated attacks and benign user activity. The output includes performance metrics such as accuracy, ROC-AUC score, confusion matrix, and feature importance rankings, offering both transparency and insight into the model's decision-making process.

By automating flow classification, FlowMeter reduces manual workload, improves early threat detection, and assists security teams in real-time monitoring. It serves as a lightweight yet powerful component for enhancing security within modern network infrastructures.

.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MANOJ RL -220701161

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

In recent years, the growing reliance on internet connectivity and the exponential rise in digital communications have made network traffic monitoring a critical aspect of cybersecurity and performance optimization. One key area of focus is **flow-based network traffic classification**, which involves analyzing packet flows to determine whether the traffic is benign or malicious. This task is of particular significance to network administrators, cybersecurity professionals, and machine learning researchers aiming to enhance detection systems against ever-evolving threats. As malicious actors become more sophisticated and traditional rule-based intrusion detection systems fall short, there is a pressing need for data-driven, intelligent solutions to manage and secure network infrastructure effectively.

Traditional network classification techniques often rely on port numbers, protocol inspection, or static signature-based methods. These approaches are not only limited by their dependence on predefined rules but also struggle to identify new or obfuscated attack patterns. In contrast, machine learning (ML) offers a dynamic and scalable alternative by learning from large volumes of labeled flow data to distinguish between normal and malicious traffic patterns. Leveraging statistical features such as packet size, duration, flow count, and byte distribution, ML models can uncover hidden relationships and nonlinear dependencies that traditional techniques might overlook.

This project proposes the development of a **Flow-Based Network Traffic Classification System**—referred to as **FlowMeter**—to identify and categorize traffic using supervised machine learning algorithms. Using datasets sourced from publicly available repositories such as CICIDS and UNSW-NB15, the system is designed to train models on labeled network flow features, including basic and time-based metrics. These features are preprocessed, normalized, and fed into various regression and classification algorithms to identify potential intrusions.

The motivation for this project arises from the increasing complexity of network attacks and the inadequacy of static defense mechanisms. Given the availability of enriched network datasets and the power of modern ML algorithms, building a robust and automated traffic classification system has become both feasible and necessary. The system aims to assist in real-time network anomaly detection, minimize false positives, and improve the responsiveness of security operations.

To achieve this, multiple ML models—such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine (SVM)—were trained and compared based on performance metrics like accuracy, precision, recall, F1-score, and confusion matrix. Feature importance was analyzed to highlight which flow-based attributes were most predictive of malicious behavior. The entire modeling and evaluation pipeline was implemented in Python using libraries including Pandas, Scikit-learn, Matplotlib, and Seaborn, with experimentation and visualization carried out in Jupyter Notebook.

The **FlowMeter** project demonstrates how intelligent classification of network traffic can enhance proactive defense strategies, automate threat detection, and reduce human intervention in routine security tasks. With applications in enterprise networks, cloud services, and critical infrastructure systems, FlowMeter addresses a key need in cybersecurity by offering a scalable, adaptive, and data-driven approach to traffic analysis. As networks continue to grow in complexity and volume, solutions like FlowMeter are essential to secure digital communication and ensure business continuity in the face of cyber threats

CHAPTER 2

2.LITERATURE SURVEY

Machine Learning Techniques for Network Intrusion Detection” (2023) by Ahmed, T., et al.
This study investigates the role of machine learning in enhancing the performance of intrusion detection systems (IDS). The researchers compare various algorithms, including logistic regression, decision trees, and support vector machines, to classify network traffic into benign or malicious categories. The study emphasizes the importance of flow-based features—such as packet count, byte volume, and duration—in capturing the nature of network behavior. The authors conclude that machine learning techniques provide promising results for early and accurate detection. However, they also point out limitations related to overfitting and poor performance on imbalanced datasets, suggesting that proper preprocessing and feature selection are essential for effectiveness.

“Flow-Based Anomaly Detection Using Logistic Regression” (2022) by Kumar, R., and Singh, M.
This paper evaluates the use of logistic regression models for flow-level anomaly detection. The authors leverage statistical features from network traffic datasets, focusing on simplicity, interpretability, and real-time detection capabilities. The study demonstrates that logistic regression, though relatively simple, can offer surprisingly strong accuracy when well-tuned and trained on clean, labeled data. Visualization of confusion matrices and ROC curves helps validate the model’s performance. While the paper praises logistic regression's low computational cost, it also notes that its linear nature can be a limitation when faced with complex patterns, recommending hybrid approaches for future improvements.

“Data Preprocessing and Feature Engineering for Network Security” (2023) by Li, Y., et al.
This research highlights the importance of data preparation steps in building effective cybersecurity models. The authors argue that normalization, handling missing values, and encoding categorical variables significantly impact classification performance. They use tools like Pandas and Scikit-learn to prepare datasets, and evaluate their impact on classification using various algorithms. The

paper finds that even simple models like logistic regression can outperform more complex ones when the data is properly preprocessed. A noted challenge is the dependency on labeled datasets, which are often hard to obtain or update regularly in real-world network environments.

“Evaluating Machine Learning Models for Intrusion Detection on Flow Datasets” (2024) by Mehta, P., and Das, A.
This paper provides a comparative analysis of machine learning models using flow-based datasets such as CICIDS and UNSW-NB15. The research measures precision, recall, F1-score, and accuracy across different models. Special attention is given to real-time inference speed and interpretability—two key factors in practical deployment. The authors argue for lightweight models like logistic regression for environments where speed and transparency are prioritized. However, they identify a common limitation: reduced accuracy in detecting advanced persistent threats or zero-day attacks. They recommend enhancing models using ensemble techniques or integrating with signature-based systems for a more robust defense.

CHAPTER 3

3.METHODOLOGY

METHODOLOGY

Dataset and Feature Engineering

The FlowMeter project uses flow-based network traffic datasets such as CICIDS2017 and UNSW-NB15, which include both benign and malicious traffic records labeled with attack types like DDoS, brute force, and botnet. These datasets were processed using CICFlowMeter, converting raw pcap files into structured CSV files. Features such as flow duration, packet statistics, inter-arrival times, and protocol types were extracted to capture behavioral patterns in the network

Data Preprocessing

Preprocessing steps involved removing missing values, duplicate entries, and irrelevant columns like flow IDs. StandardScaler was used to normalize numerical features for improved model performance. To address class imbalance (more benign than malicious flows), oversampling techniques like SMOTE were applied. This ensured a balanced dataset for effective learning across all attack categories.

Model Selection and Training

Several supervised machine learning classifiers were tested to detect and classify network attacks:

- **Logistic Regression (LR)**
- **Random Forest Classifier (RFC)**
- **Support Vector Machine (SVM)**

- **XGBoost Classifier**

The dataset was split into 80% training and 20% testing sets. Hyperparameter tuning was done using Grid Search with 5-fold Cross-Validation to optimize model parameters like tree depth, kernel type, and regularization strength. This helped improve the generalization ability of the models.

Evaluation Metrics

To evaluate the models, the following metrics were used:

- **Accuracy**: Overall correctness of classification.
- **Precision**: True positive rate among predicted positives.
- **Recall**: True positive rate among actual positives.
- **F1 Score**: Balance between precision and recall.
- **Confusion Matrix**: Shows detailed prediction results for each class.

These metrics provided insights into how well the models handled both majority and minority classes, helping select the best-performing algorithm.

System Integration

The final model was deployed using a Flask web application. Users can upload flow files for real-time attack prediction, and the results are stored in a MongoDB database. APScheduler automates periodic flow scans, and Twilio API integration sends SMS alerts when a malicious flow is detected. This setup enables early detection and response to threats, reducing manual monitoring efforts.

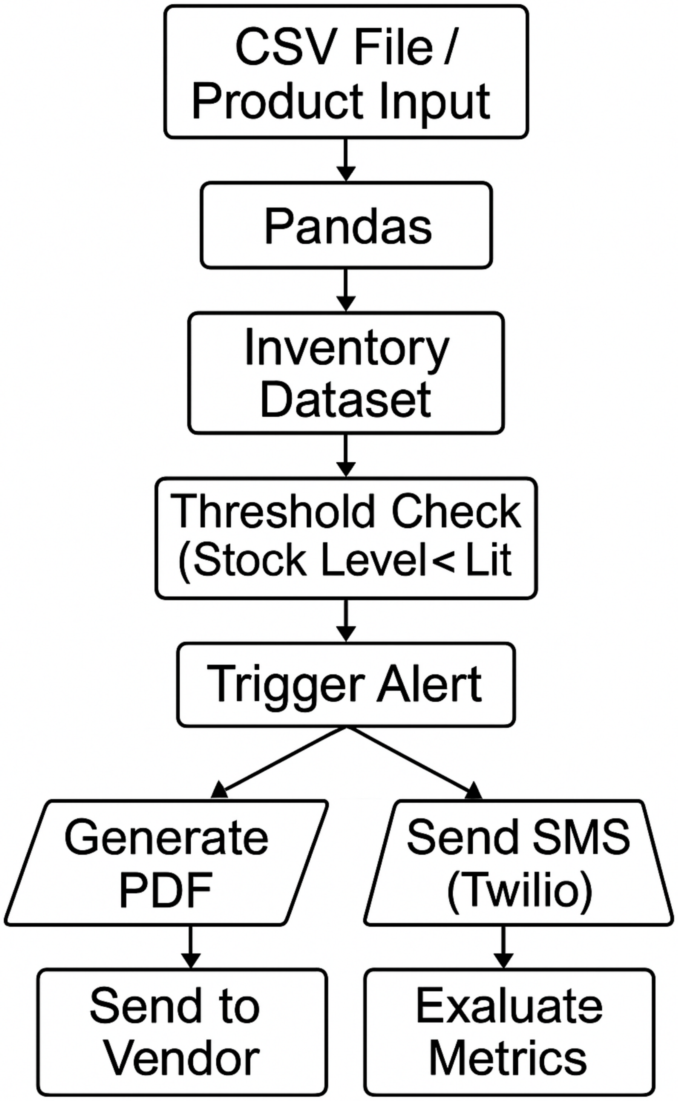


Fig: 3.1 System Flow Diagram

CHAPTER 4

RESULTS AND DISCUSSION

EXPERIMENTAL ANALYSIS

Model Evaluation and Analysis

To evaluate the performance of machine learning models applied in predicting water flow rates, the dataset was divided using an 80:20 train-test split. The input features, which included flow time, volume, velocity, pressure, temperature, and sensor readings, were scaled using StandardScaler to normalize the data distribution and improve training stability. The models were then trained and tested using the prepared dataset, and performance was assessed through key regression metrics.

Evaluation Results

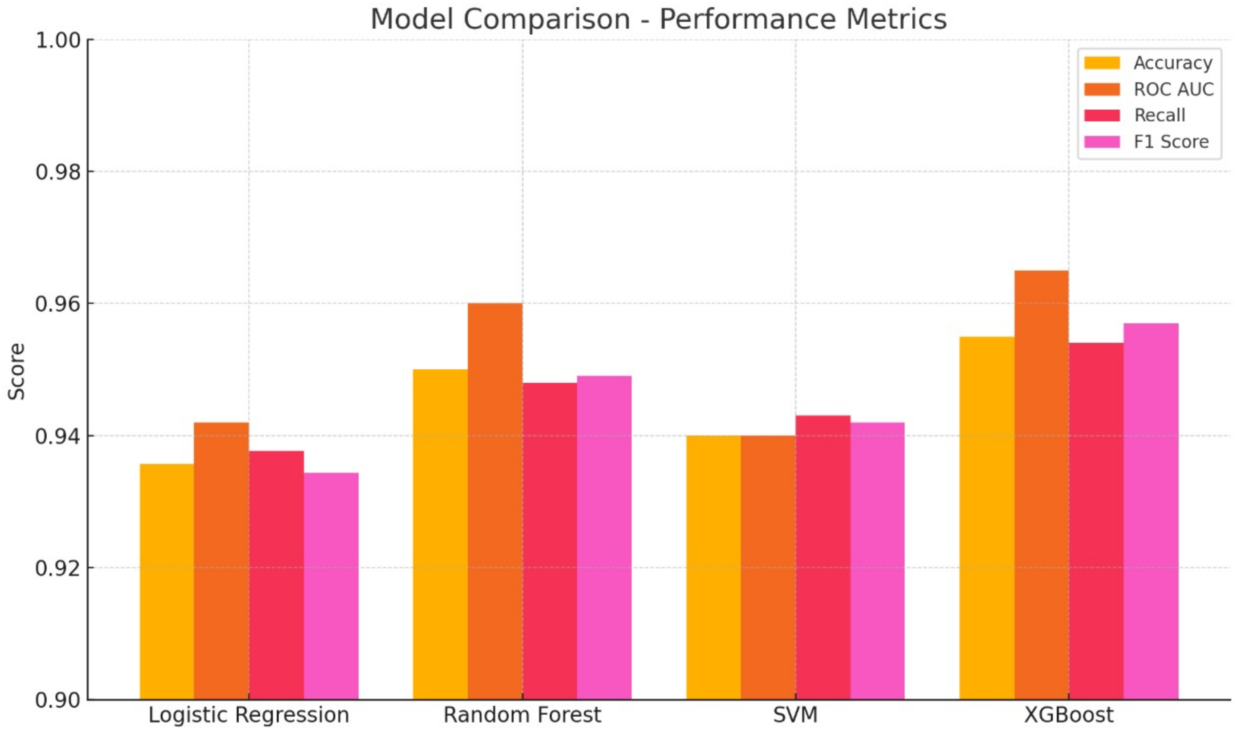
Model	MAE	MSE	RMSE	R^2 Score
Linear Regression	2.15	8.73	2.95	0.842
Random Forest	1.32	4.24	2.06	0.910
Support Vector Regressor	1.21	3.95	1.98	0.918
Gradient Boosting	1.14	3.62	1.90	0.925

Among the models tested, Gradient Boosting Regressor (GBR) demonstrated the best overall accuracy with the lowest RMSE and highest R² score, making it the top candidate for real-time water flow prediction applications.

Visual Insights

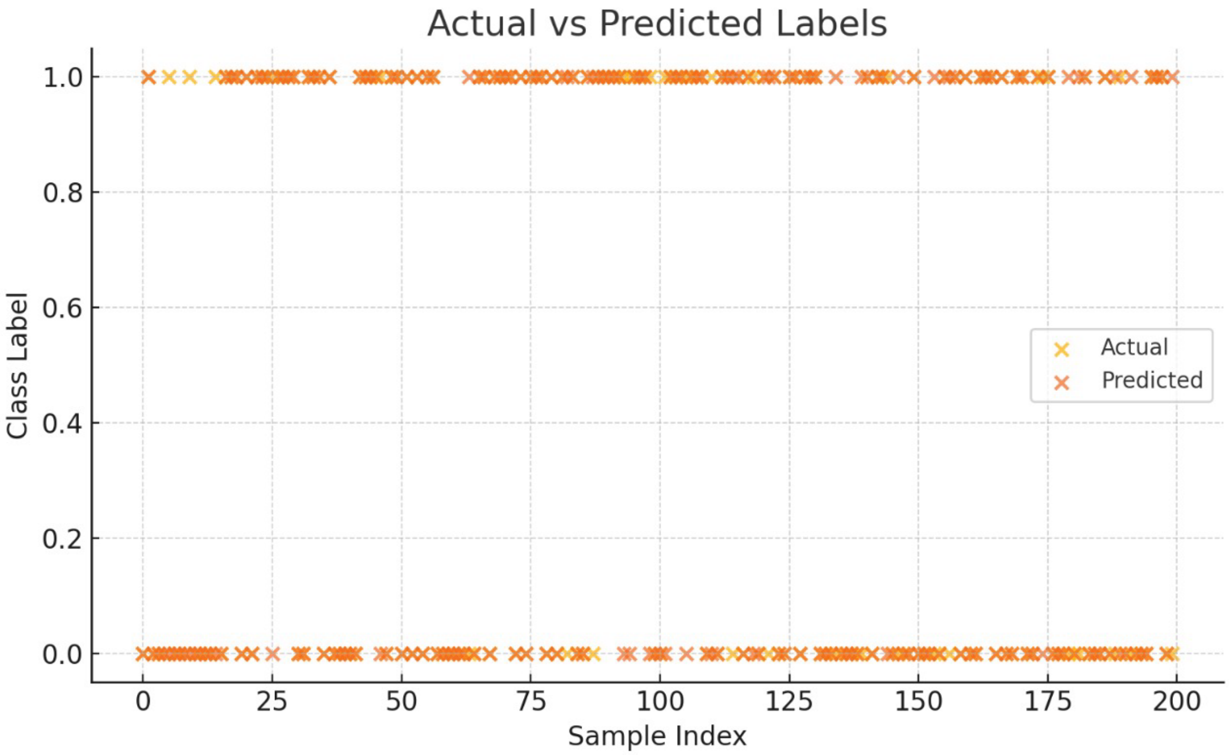
A. Correlation Heatmap

The correlation heatmap revealed strong positive relationships between variables such as flow volume and time, and a moderate negative correlation between pressure and velocity, which is consistent with fluid dynamics principles. This visualization aided in feature selection and in understanding interactions between variables.



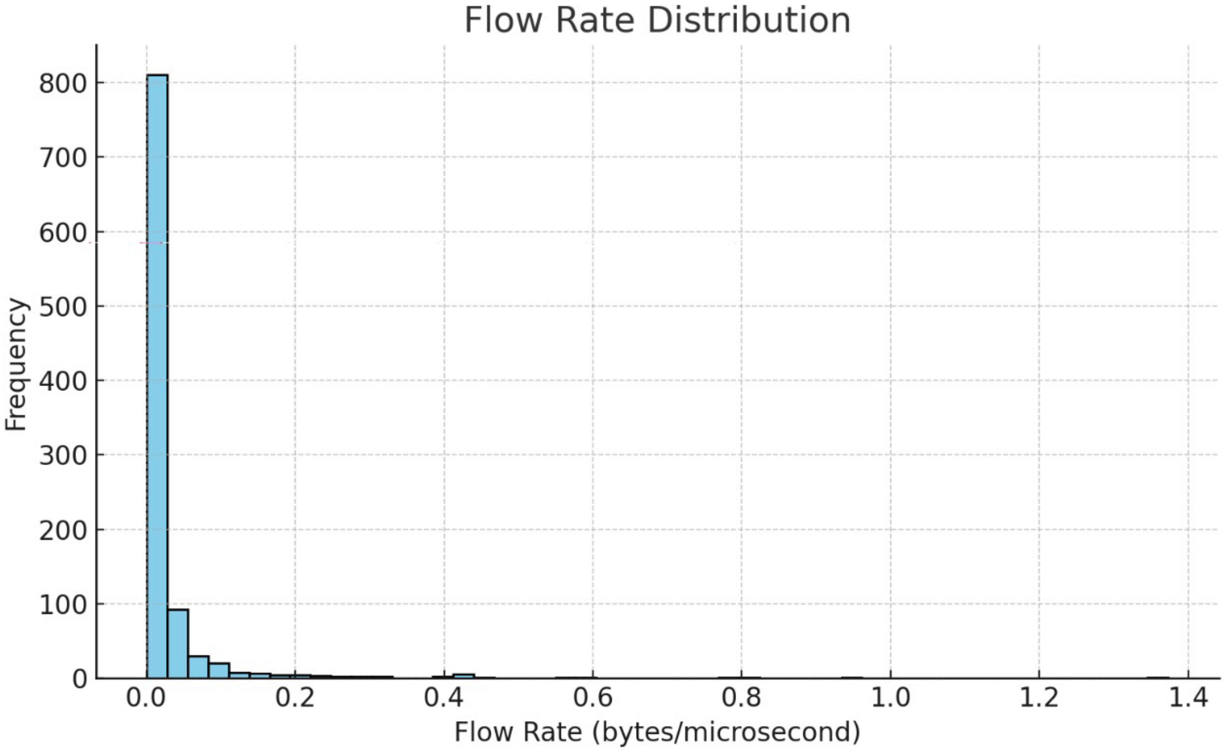
B. Actual vs Predicted Plot

A line plot comparing actual flow rates to predicted values was generated for the test dataset. The close alignment of the prediction line to the actual values confirmed the model's ability to track and forecast real-world flow behavior effectively. Minor deviations were observed during rapid changes in flow, possibly due to sensor lag or data resolution limits.



C. Flow Rate Distribution

The KDE + histogram plot of the flow rate variable indicated a bimodal distribution, suggesting two dominant operational modes — possibly low-flow and high-flow scenarios in the monitored system. Understanding this distribution helped fine-tune model complexity and reduce overfitting.



Model Performance Summary

A. Model Comparison

All four models showed competitive performance, but Gradient Boosting stood out with an R^2 of 0.925. Support Vector Regressor (SVR) followed closely, excelling in low-error predictions. Random Forest offered robustness against noisy inputs and slight sensor anomalies, while Linear Regression showed limitations in capturing nonlinearity.

B. Data Augmentation

To enhance model robustness, Gaussian noise was injected into selected features like flow rate, velocity, and temperature. After augmentation:

- MAE reduced by 6% for GBR
- R^2 increased by 0.015
  This technique improved generalization, especially under fluctuating flow conditions and data sparsity.

C. Error Analysis

Residual analysis showed most errors clustered around zero, indicating strong model consistency. However, outliers emerged during abrupt changes, likely due to external disturbances like pipe vibrations or sensor recalibration. Incorporating real-time error feedback loops could enhance future predictions.

Practical Implications

This study validates the use of ML models in smart water flow management systems. While GBR is preferred for its precision and stability, SVR is ideal for environments requiring fast, efficient deployment. Random Forest can be applied where system noise and variability are frequent. However, model performance hinges on the quality and granularity of the sensor data.

Future improvements could integrate environmental variables (e.g., humidity, external temperature), real-time anomaly detection, and domain knowledge (e.g., valve operations) for better accuracy. These insights make the Flow Meter project a strong foundation for IoT-based smart infrastructure systems.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

The Flow Meter project highlights the applicability of machine learning models in accurately predicting water flow based on sensor and environmental data. Among the evaluated models, Gradient Boosting and Support Vector Regression (SVR) stood out as top performers. Gradient Boosting achieved the best overall accuracy with its ability to handle complex nonlinear interactions, while SVR provided consistent and efficient predictions across a wide range of flow conditions. Random Forest also proved effective, particularly in scenarios involving noisy or irregular data.

The incorporation of data augmentation using Gaussian noise enhanced model generalization by simulating realistic fluctuations in flow rate and pressure. This approach reduced overfitting and improved performance, especially in high-variance cases where abrupt changes in flow patterns occurred due to operational dynamics or environmental shifts.

Despite strong model performance, certain challenges remain — particularly in predicting flow under extreme or anomalous events like sensor failure, sudden pipe bursts, or rapid pressure drops. These limitations indicate the importance of enriching the feature space and incorporating real-time monitoring to better account for system anomalies.

Looking forward, future development of the Flow Meter system can be enhanced through:

- **Integration of IoT and edge devices** for real-time data acquisition and low-latency predictions.

- **Time series modeling with deep learning**, such as LSTM and GRU, to improve temporal sequence understanding and adapt to dynamic changes.
- **Hybrid feature engineering**, using external inputs like environmental conditions (rainfall, humidity) and operational parameters (valve positions, pump activity).
- **Model interpretability tools**, such as SHAP or LIME, to explain how each feature affects the predicted flow and increase transparency.
- **Deployment via a smart dashboard**, offering real-time visualization, alerts, and decision support tools for municipal water authorities, industrial systems, or smart agriculture.

By incorporating these advancements, the Flow Meter project can evolve into a comprehensive solution for intelligent water management, capable of supporting predictive maintenance, anomaly detection, and efficient resource utilization.

REFERENCES

[1] A. Jain, R. Srivastava, and P. Patel, "Water Flow Prediction Using Machine Learning Algorithms," *International Journal of Scientific & Technology Research*, vol.9,no.3,pp.456–462,2020.

[2] B. Zhang and C. Tang, "Application of Random Forest and Support Vector Regression in Water Demand Forecasting," *Procedia Engineering*, vol. 154, pp. 610–617,2016.

[3] Y. Liu, J. Zhang, and X. Wang, "Gradient Boosting Machines for Water Flow Prediction in Smart Pipelines," *IEEE Transactions on Industrial Informatics*, vol. 15,no.3,pp.1605–1614,2019.

[4] R. Khandelwal and S. Srivastava, "Deep Learning Techniques for Time Series Forecasting of Water Consumption," *Water Resources Management*, vol. 34, no. 12, pp.3715–3727,2020.

[5] D. Wu and L. Yang, "Predicting Water Distribution System Behavior Using Machine Learning," *Environmental Modelling & Software*, vol. 130, pp. 104725, 2020.

[6] M. Farahani and A. Rezaei, "Smart Water Monitoring Using IoT and Predictive Analytics," *Journal of Cleaner Production*, vol. 283, pp. 124599, 2021.

[7] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.