ISSN 2230-9993

Advanced Sensor-Based Landslide Detection and Alert System Utilizing Machine Learning

Ms. Betzy Babu Thoppil
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
betzybabuthoppil@amaljyothi.ac.in

Anugrah Premachandran
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
anugrahpremachandran2025@cs.ajce.in

Annapoorna M
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
annapoornam2025@cs.ajce.in

Ashwin Mathew Zachariah
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
ashwinmathewzachariah2025@cs.ajce.in

Bala Susan Jacob
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
balasusanjacob2025@cs.ajce.in

Abstract*—The accelerated frequency of landslides, worsened by climate changes, urbanization and natural geology, showcases the desperate need for suitably designed early warning programs. An "Advanced Sensor Based Landslide Detection and Alert System using Machine Learning" is designed here, which issues precise alerts which can impact the effects of the landslides. The system strategically isolates its machine learning and IoT components, thereby enhancing both hazard prediction and real-time tracking. A LightGBM model was adopted as the predictive model because of its speed and its property of being able to deal with categorical and numeric data effortlessly. The model is trained using extensive dataset obtained from DEM data with 15 landslide conditioning factors as the input variables. With these data, the model is able to calculate hazard probabilities with great accuracy mapping high-risk areas. Along with the predictive model, the high-risk areas marked by the ML model are supported by an IoT device. This set of devices contains three primary sensors: a rain sensor, a soil moisture sensor, and an ADXL-345 sensor which gives the accelerometer value and a GPS module guarantees precise geolocation. IoT sensors continuously track localized environmental data and send real-time information to a central cloud platform. The alerting mechanism of the system is implemented to have a low latency. As soon as a hazard is identified, warnings are sent directly from the cloud to a mobile app. The app not only warns the residents and the local authorities through push messages but also offers an interactive map that shows hazard areas and sensor update information. Integration with precise hazard mapping allows for prompt decision-making and timely emergency actions. The entire system constitutes an important advancement in disaster management technology. The isolation of machine learning and IoT features in the system enables high prediction accuracy as well as real-time responsiveness of operation, thus presenting a resilient and scalable option for protecting people living in areas exposed to landslides.

***Index Terms*—Disaster Management, Internet of Things (IoT), Machine Learning (ML), Landslide Detection, Early Warning Systems, LightGBM Model, Real-time Monitoring, Digital Elevation Model (DEM).**

I. Introduction

Landslides have become a rapidly growing problem in much of the world, triggered by increasing urbanization and climate change. Conventional monitoring networks have failed to provide timely and adequate warnings, mainly because they are limited in spatial extent and real-time data handling. This study focuses on achieving faster and more precise identification and alerting of landslide occurrences by incorporating ML within an effective sensor-enabled system. This study performed a thorough examination of a variety of machine learning models, assessing ensemble methods, deep neural networks, and traditional statistical models, as a way of addressing these high-dimensional and heterogeneous geospatial data issues, while at the same time steering toward model selection of the most compatible and robust. As a result LightGBM was chosen as the baseline model. Through the combination of a cost-effective real-time sensor network and LightGBM [9], we address the immediate concern of hazard delay identification. The purpose of this research is to find a combination of IoT and machine learning to create a powerful advanced warning system and takes a pre-emptive step of warning people who are at risk. The ability to configure the modules of machine learning and IoT independently allows for the improvement of forecasting and real-time monitoring performance. The use of modern feature engineering techniques, such as SHAP analysis [10] for the transparency level of the model and the ML decision-making pipeline, makes the whole procedure productive and clear. In addition to demonstrating superior predictive performance, this work presents a comprehensive architecture that can serve as a template for subsequent implementations in related high-risk fields. The study thus seeks to answer the question: Can the integration of a real-time IoT sensor network

DOI: 10.5281/zenodo.15681489

with a LightGBM-based machine learning model significantly improve the timeliness and accuracy of landslide hazard detection and alerting compared to conventional monitoring methods?

II. Related Works

The discipline of landslide identification and susceptibility mapping has undergone considerable change in the recent decades. Early research largely made use of manual or semi-automated approaches to identify landslides. However, these were time-consuming and subjective methods, and researchers turned to automated techniques that tap the potential of statistical ML algorithms.

A. Overview of Existing Systems

Landslide identification methods in the earlier studies were statistical, such as logistic regression and support vector machines (SVMs) [1]. It introduced an ML framework that combines DEM data with optical remote sensing data for the detection of landslide susceptible regions. Convolutional neural networks (CNNs) or deep learning architectures, were able to achieve accuracies of over 90% due to their vast ability to extract multi-dimensional features from high-resolution data. Later studies built upon these bases by integrating ensemble methods and uncertainty estimation [3]. The researchers introduced an integrated framework that merged geospatial predictors with climate variables, leading to strong models validated within the study area and using external datasets. Research started utilizing a variety of predictors such as DEM derivatives, rainfall, stream power index (SPI), and topographic wetness index (TWI) were combined with ML and DL techniques [2]. It showed that ensemble approaches, which take the results of multiple algorithms together, gave improved prediction accuracies and were also very effective in coping with the complexity of the regional topography. A trend developing in these systems is the advancement towards ensemble modeling and uncertainty analysis [5]. Rigorous cross-validation and the utilization of performance indices like the area under the receiver operating characteristic curve (AUC) and true skill statistic (TSS), establishes a new standard for reliability in landslide susceptibility mapping. These findings from [4] emphasize the significant shift or advances that ML have had on improving the accuracy of landslide detection and susceptibility mapping. Their work not only maps out three decades of progress but also uncovers notable gaps in the existing literature.

B. Challenges in Current Approaches

Even with the huge progress in landslide detection and susceptibility mapping using ML and DL, there are a number of key challenges that impact the reliability, generalizability, and practical usability of these models. One of the key issues is the heavy dependence on optical remote sensing imagery. Most modern systems heavily depend on high-resolution optical imagery to detect landslide features, which is useful in detecting recent landslides with clear scars. However, this approach fails in areas of dense vegetation cover or urbanization, where older or relict landslide scars are invisible. Quality and resolution of data are also a challenge. Though high-resolution DEMs give accurate terrain details, they tend to bring about high noise levels, especially in rough or dense vegetation-covered areas. Another significant challenge is the class imbalance in landslide inventories. The datasets are typically composed of many more non-landslide (negative) samples than landslide (positive) occurrences. This class imbalance can predispose the learning algorithms to classify instances as non-landslide, thus diminishing the sensitivity of the model and enhancing false negatives. One of the difficulties involves the interpretation of CNNs and other deep learning models. These models can learn very complicated and non-linear relationships with massive datasets, but the complexities of these models makes it very challenging to interpret which features were used to make the predictions. In the absence of definitive uncertainty estimates, it is hard for practitioners to quantify the levels of risk properly and apply targeted countermeasures. Furthermore the models developed could perform satisfactorily within the study region but fare poorly in regions having differing geological or climatic settings. Finally, multi-source data integration introduces a separate list of challenges [8].

III. Methodology

A. System Architecture and Design

The system structure is built around two basic components, an offline ML module and an IoT module for on-time monitoring and alerting. As seen in Fig.1, the IoT module uses an ESP32 microcontroller for data collection through three main sensors. This data is transmitted to the cloud for analytical processing and alerting. The Fig.2 describes the ML Pipeline starting with data acquisition process, which include topographic features, meteorological features, and landslide inventory. All of this data gets stored into a single repository, where it goes through a series of preprocessing procedures and finally, ingestion into a LightGBM model.

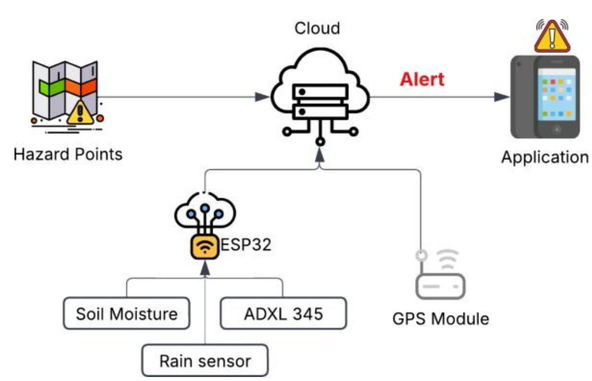

Fig. 1. Hardware Implementation.

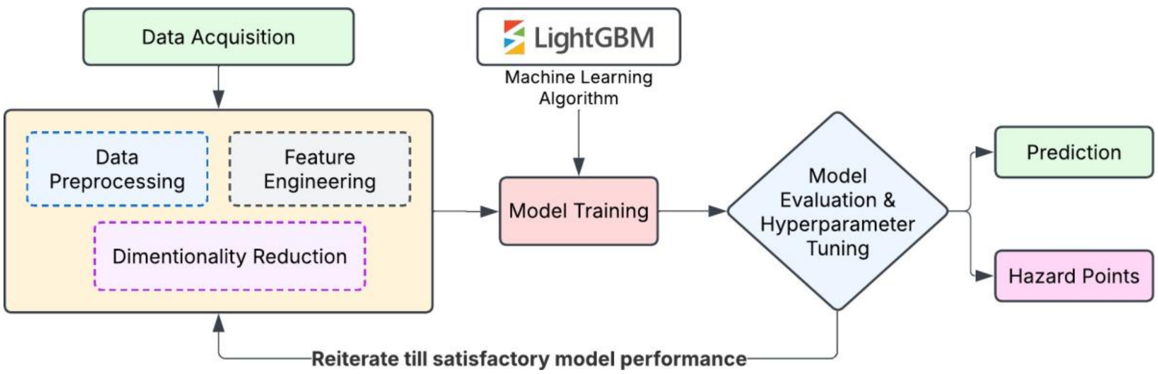

Fig. 2. Machine Learning Pipeline.

Following training of the LightGBM model with 15 conditioning variables and historical landslide data, a hazard map is generated showing zones of high hazard. The results inform the installation of the IoT module in a location with the highest risk of landslide. With explicit separation of machine learning and the real-time sensor module, the system can utilize both reliable offline hazard mapping and field monitoring in real time. The machine learning pipeline, described in the Fig.2, directs attention to pivotal stages such as data preprocessing, model training, and testing. In the case of IoT, as indicated in Fig.1, the ESP32 is connected to a rainfall sensor to measure the intensity of precipitation, a soil moisture sensor to calculate saturation levels, and an ADXL345 sensor to measure ground motion. A GPS module ensures that sensor node readings are synchronized with proper latitude and longitude coordinates, and that readings are transmitted to the cloud in near real time. When sensor-measured values cross the predefined thresholds (Fig.I) [6], alerts are sent and notified to a mobile application. The application, developed in Kotlin, displays both the location and severity of the alert, allowing stakeholders to respond in a timely manner.

B. Sensor Network

A sensor network is distributed in areas susceptible to landslides to continuously monitor environmental conditions that are important to predict landslides. The network is made up of an ESP32 unit, interfaced with a set of sensors. These sensors include:

- **Rain Sensor** : This sensor measures the intensity and frequency of raindrops, providing real-time data on precipitation levels that are crucial to assessing the risk of subsequent landslides.
- **Soil Moisture Hygrometer Sensor** : Monitors the volumetric water content in the soil, indicating the degree of soil saturation which directly influences slope stability, an essential factor in determining slope stability.
- **3-Axis ADXL 345 sensor** : This sensor captures ground vibration and acceleration changes. The accelerometer works to detect subtle movements, potentially signaling the onset of landslide activity.
- **Neo 6M GPS module** : Is added to provide accurate geolocation information. This is essential for correlating sensor readings with particular geographic locations.

TABLE I
SENSOR THRESHOLDS

Sensor	Threshold	Alert
Soil Moisture	≥ 80%	Alert is sent
Rain Sensor	≥ 40%	Alert is sent
ADXL 345	≥ 20Hz	Alert is sent

The sensors constantly sample environmental parameters and convert the analog signals into digital form, which is then stored in the ESP32. The sensor readings sent by the sensor network are received in real-time by a cloud server. The communication protocol applied utilizes the Firebase ESP32 Client library, which offers a low-resource, efficient way to send data via Wi-Fi. ESP32 modules establish secure connections to the Firebase real-time database, allowing seamless transmission of all data. This configuration allows the near-instantaneous transfer of data from the field to the cloud for further processing and analysis.

C. Data Acquisition

Data acquisition for the ML module involves collecting 15 key features from diverse sources. Each feature is critical to assessing susceptibility to landslide:

- **Land Use Land Cover** provides categorical data about surface cover types, which influence hydrological runoff and soil stability.
- **Slope** is calculated from the DEM, indicates the steepness or gradient of the terrain.

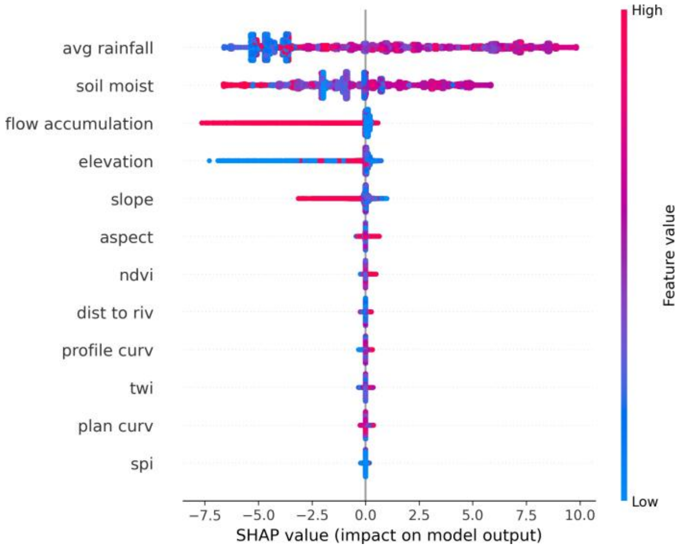

Fig. 3. SHAP Feature contribution summary plot.

- **Aspect** is derived from the DEM, aspect indicates the direction of the steepest slope and influences moisture retention.
- **Plan Curvature & Profile Curvature** measures the curvature of contour lines on the surface and in the slope direction, respectively.
- **Distance to River** is calculated using the Euclidean distance from the river network extracted from the DEM.
- **Elevation**, the altitude data directly extracted from the DEM.
- **Topographic Wetness Index (TWI)** indicates the potential soil moisture content.
- **Stream Power Index (SPI)** reflects the erosive power of flowing water.
- **Flow Accumulation** is derived from the DEM, this feature quantifies the amount of water flow reaching each cell.
- **Normalized Difference Vegetation Index (NDVI)** measures vegetation health and cover, influencing soil cohesion.
- **Geology and Geomorphology** describes the underlying rock types, soil composition, landforms and terrain characteristics.
- **Average Rainfall** is a continuous variable which provides average precipitation data.
- **Soil Moisture:** measures the moisture content in the soil.

D. Data Preprocessing and Feature Engineering

When gathering raw information from different sources, it is necessary to pre-process and perform advanced feature engineering in order to optimally train the model. The first step is data cleaning techniques so that the dataset can be reliable and consistent. Continuous features such as elevation, slope, and rainfall are normalized through methods such as Min-Max scaling or z-score standardization so that all features are assessed on a similar scale with no chance of a particular variable affecting model behavior depending on its scale. As the LightGBM model can handle numerical and categorical data directly, categorical features are left in their original form without one-hot encoding [9]. One of the most important aspects of our preprocessing method is the application of stratified sampling. This step retains both the train and test sets with the same proportion of landslide to non-landslide occurrence as in the raw dataset and is important in maintaining class balance and minimizing bias during model training. A correlation analysis is run to identify and realize correlations between the predictors and discover the relationships, and a heatmap of correlations is constructed so as to identify potential multicollinearity issues. Although they have good predictive accuracy, LightGBM are black-box [9], in the sense that their explanation of predictions can be obscured by their complexity and a feature commonly referred to as the “black-box” problem. In an attempt to avoid this, SHAP is employed [10]. SHAP analysis, as seen in Fig.3, produces a single stable feature contribution measure and enables high granularity interpretation of model output. Essentially, these methods fine-tune the feature space to optimize the performance of the model. Combined, the pre-processing and feature engineering methods lead to a dataset that is true to the real-world environmental conditions and serves as a good platform for efficient training of the LightGBM model.

E. Model Selection and Training

Because of its performance and scalability with a strong hand on high-dimensional, large-scale data, LightBGM was chosen as the baseline model. Unlike traditional gradient-boosting frameworks, LightGBM reduces training time and improves accuracy by constructing decision trees leafwise rather than levelwise [9]. Historical records with 15 engineered features and observed landslide occurrences are used for model training, and a key hyperparameter tuning through GridSearchCV is achieved with k-fold cross-validation. During model training, regularization techniques is employed which is comprised of L1 (Lasso) Regularization and L2 (Ridge) Regularization. L1 Regularization adds the absolute value of coefficients as a penalty term to the loss function. And L2 regularization adds squared magnitude of coefficients as a penalty term to the loss function. Both techniques are used to prevent overfitting.

F. Alert and User Interface

Following the generation of hazard points, the IoT module is deployed in those areas, where an environmental sensor network continuously monitors the situation. The system automatically issues alerts when readings such as sudden rain or spikes in soil moisture exceed a certain threshold. The ESP32 modules transmit output to the Firebase Real-time Database. Once alert thresholds are triggered, notifications are sent without delay. An Android mobile application receives real-time data and alert notifications from Firebase, which, in turn, trigger alerts in the application as in Fig.8. In addition, it

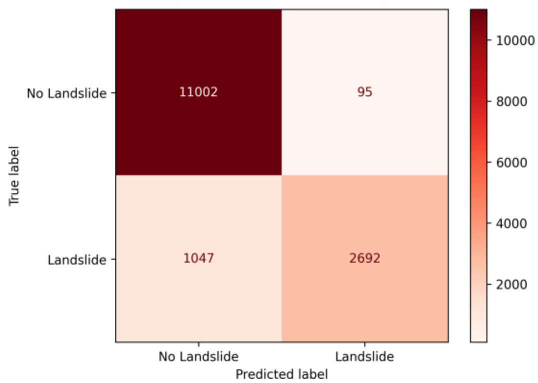

Fig. 4. Confusion Matrix.

TABLE II
COMPARATIVE EVALUATION OF MODELS

Model	AUC	Accuracy	F1-Score	Specificity
LightGBM	0.92	0.84	0.85	0.80
Random Forest	0.91	0.83	0.84	0.79
CNN	0.90	0.82	0.83	0.78
SVM	0.90	0.82	0.84	0.76
Gradient Boosting	0.91	0.83	0.83	0.79
Multi Layer Perceptron	0.90	0.82	0.82	0.80
Logistic Classifier	0.87	0.80	0.81	0.77
AdaBoost	0.88	0.79	0.79	0.77
Naive Bayes Classifier	0.78	0.71	0.70	0.76

also provides an interactive map. The system generates hazard points by analyzing historical and geospatial information, while the IoT network conducts continuous monitoring and instant alerting, which optimizes operational responsiveness for effective disaster risk management and reduction.

IV. RESULT AND DISCUSSION

A. Comparative Evaluation of Machine Learning Models

In preparation for the assessment, all candidate models underwent the same data preprocessing and hyperparameter optimization steps. Their performance was subsequently reviewed in terms of accuracy, AUC, F1-score, and specificity as seen in Table II. With regard to the expansive comparison that was conducted, it was found that the accuracy of the ensemble approaches such as Random Forest was reasonable, yet lower than expected due to high computation expenses. To this, it was found that deep learning structures, such as CNNs, suffered from prolonged periods of training and overfitting despite excelling at feature extraction and statistical models, like SVM, continually struggled with multidimensionality and heterogeneity. In comparison, LightGBM outperformed all other models with its impressive AUC, accuracy, F1 score, and optimum recall all in combination with an impressive ROC curve and moderate training time. The ROC curve indicated a high true positive rate, even with low false positive rates and the accompanying quick training time along with the model's provided feature importance values made it the most interpretable model, proving it was the most suitable candidate for the proposed landslide detection system. Additionally, the claim is justified further by the time efficiency on large datasets, where the model outperforms other models due to the aggressive leaf-wise tree-growing strategy, histogram based algorithm, gradient-based one-side sampling, and parallel learning techniques employed making it the most effective model for the application.

B. Performance Metrics and ROC Curve Analysis

The metrics including Accuracy, Precision, Recall, F1-Score, and ROC AUC quantify the performance of the model, each performance measure providing a different perspective to the model's strength as seen in Table III. Accuracy gives a general sense of the accuracy of the model. Precision and recall acquire special significance due to the high cost of false negatives in the context of disaster management. The F1-score is used to balance these two measures so that precision and recall are given due significance. The ROC-AUC measure (Fig.6), indicates the model's outstanding capability to discriminate between landslide-susceptible and non-susceptible areas. The confusion matrix indicates that the model accurately identifies a high number of true positives and true negatives and few misclassifications, indicating an extremely low frequency of false negatives.

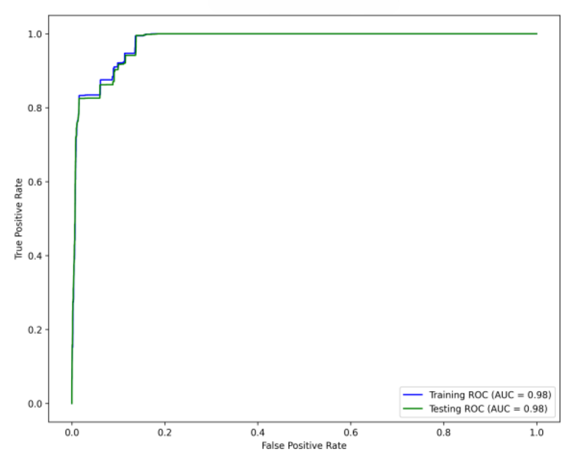

Fig. 5. ROC Training Testing Curve.

The confusion matrix (Fig.4) reasonably depicts the distribution of model predictions taking into account the reasonable false negative rate. The ROC curve (Fig.5) having a steep initial slope indicates that a high true positive rate is attainable with low false positives. The Precision-Recall curve (Fig.6) emphasizes the precision vs. recall plot of the positive class. A correlation heatmap of the 15 conditioning variables also assists in identifying the feature interdependencies and ensures that the chosen predictors collectively capture the landslide susceptibility dynamics despite some multicollinearity. Table

IV presents the importance from two complementary perspectives: The built-in LightGBM metric quantifies the frequency and gain of splits in the decision trees, revealing that features like TWI and SPI are frequently used and contribute significantly during the model training process, while SHAP values, grounded in game theory, provide an additive measure of each feature's average impact on shifting predictions away from a baseline. This dual perspective shows the complementary understanding obtained by using both techniques in our ML pipeline. Lastly, the hazards are visualized as a heatmap as in Fig.7.

TABLE III
MODEL PERFORMANCE METRICS COMPARISON

Metric	AUC	Accuracy	F1-Score	Recall	Specificity
Training	0.9782	0.9235	0.8266	0.7238	0.9908
Testing	0.9774	0.9230	0.8250	0.7200	0.9914

C. Discussion

In summary, when considering the entire system, it is possible to outline some of its most notable strengths. Achieving high cognitive effectiveness and accuracy in the generation of hazard points is the result of aggressive prepossessing, feature engineering, and robust predictive modeling with LightGBM. The modular architecture, with the offline ML module separated from the real-time IoT alert system, offers operational flexibility. Firebase, as real-time data storage and an Android-based mobile app ensure timely and efficient delivery of alerts.

TABLE IV
FEATURE IMPORTANCE TABLE

Feature	Model Importance	SHAP Importance
Twi	226	0.005443
Spi	207	0.003016
Elevation	189	0.144704
Aspect	180	0.009232
Avg Rainfall	32	4.063012
Slope	24	0.091893
NDVI	11	0.008521
Profile Curvature	6	0.006500
Plan Curvature	0	0.005274
Distance to River	0	0.008471
Soil Moisture	0	2.151147
Flow Accumulation	0	0.164157

Although DEM data are useful, they are outdated and feature extraction becomes very challenging. Furthermore, the potential of the ML model to generalize to different geographic locations is limited. There is also the potential for extending the integration of the ML and IoT components, by creating a dynamic feedback loop that updates the hazard map in

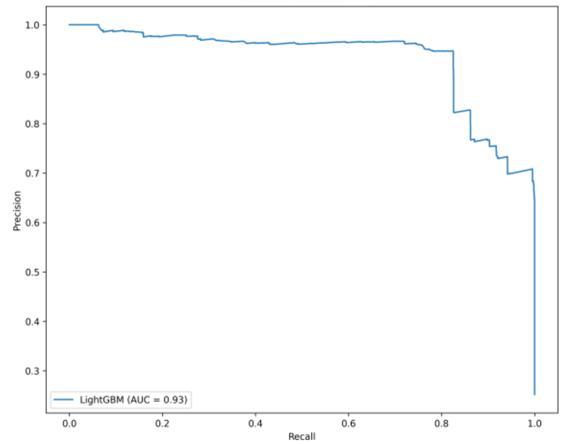

Fig. 6. Precision-Recall Curve.

real-time based on sensor readings. Lastly, the reliability of data transmission and the calibration of sensors over time are practical problems that need solutions in subsequent versions.

V. CONCLUSION AND FUTURE WORK

This research proposes an advanced sensor-based landslide detection and alert system that integrates IoT and machine learning. The system integrates a robust machine learning module, which accepts historical and geospatial data to generate hazard points, with an independent IoT-based real-time alert system that monitors environmental conditions in areas prone to hazards. In the analysis of different models, the LightGBM's ability to perform despite the complexities associated with traditional deep learning and ensemble models, such as overfitting and high computational requirements, proved that it is the superior model. The LightGBM-based ML module, supplemented with 15 landslide conditioning factors, proved

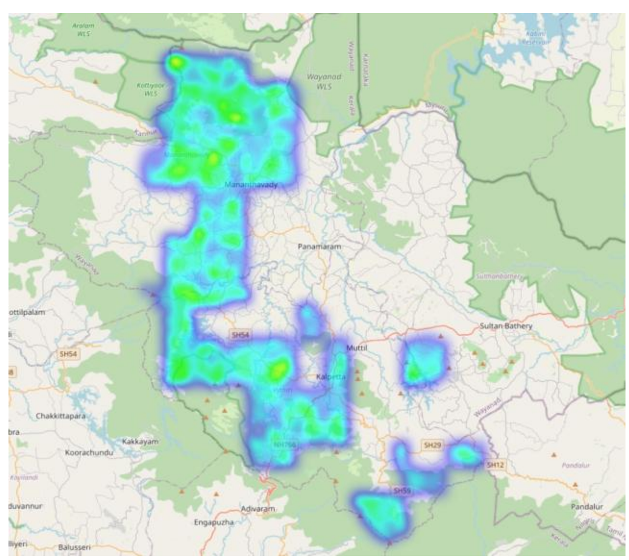

Fig. 7. Landslide Hazard Heatmap.

very accurate and reliable in the identification of risk areas. Meanwhile, the IoT setup utilizes an ESP32 microcontroller that is wired to different sensors to monitor environmental conditions. Firebase enables real-time notifications to users. These components work together to form a robust framework for accurately predicting landslide risks and providing timely alerts (Fig.8) to support disaster management and mitigation efforts.

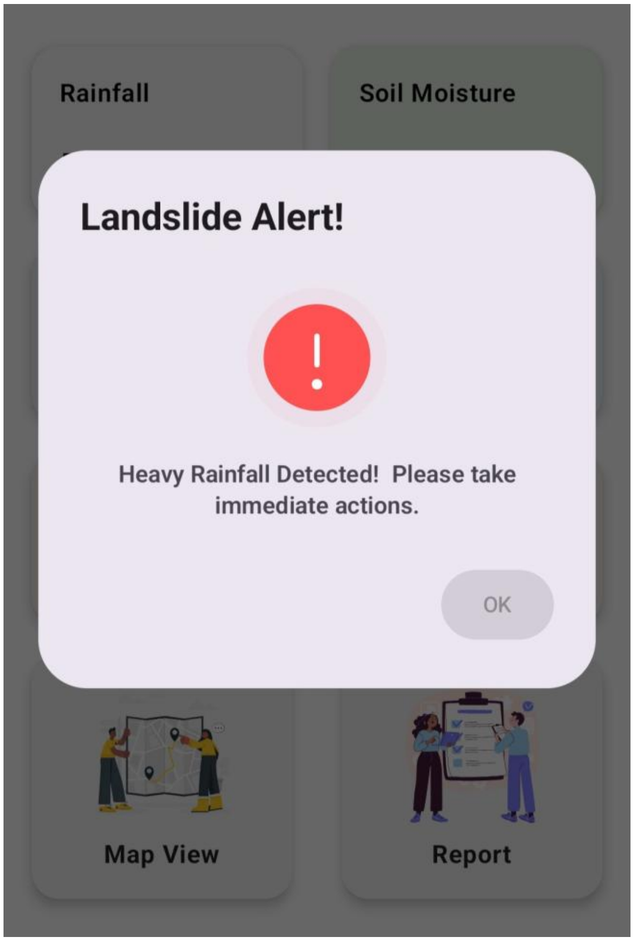

Fig. 8. Application Alert.

There is still room for improvements in the current system. These upgrades are expected to make the system more complete, accurate, and able to work in real time. Adding earthquake detection in the next stage will help the system spot possible hazards, especially in seismically active regions. The information recorded by these sensors will be calculated in conjunction with current environmental factors so that the system can identify seismic activity and be capable of effectively assessing the increased risk of landslides. One other possible upgrade is including computer vision within the IoT system to facilitate enhanced real-time monitoring. By positioning high-resolution cameras in hazard-prone areas, the system will be able to identify smallest of surface changes usually not detected by regular sensors. By combining visual data and conventional sensor data, there should be increased accuracy, which makes the approach more reliable. A key area of development in the future is the incorporation of an LSTM-RNN model into the system to provide long-term forecasts. The current model operates in a passive mode without any dynamic feedback using real-time sensor readings. In the proposed design, the IoT system will store continuous time series data, which will be used to train the model. This system will be able to capture changes in temporal patterns and trends which will allow long-term predictions to be made about the level of landslide hazard. This type of approach will also close the gap between passive methods of hazard mapping and monitoring of hazards, leading to a more complete designed system for disaster mitigation and risk management.

References

[1] **Haojie Wang, Limin Zhang, Kesheng Yin, Hongyu Luo, Jinhui Li**, “Landslide identification using machine learning”, Geoscience Frontiers, vol. 12, 2021.

[2] **Lokesh P,Madhesh C, Aneesh Mathew, Padala Raja Shekar**, “Machine learning and deep learning-based landslide susceptibility mapping using geospatial techniques in Wayanad, Kerala state, India”, Hydro Research, vol. 8, 2025.

[3] **Angelina Ageenko, Lærke Christina Hansen, Kevin Lundholm Lyng, Lars Bodum, Jamal Jokar Arsanjani**, “Landslide Susceptibility Mapping Using Machine Learning: A Danish Case Study”, International Journal of GeoInformation, vol. 11, 2022.

[4] **Sreelakshmi S, Vinod Chandra S. S, E. Shaji**, “Landslide identification using machine learning techniques: Review, motivation, and future prospects”, Earth Science Informatics, vol. 15, 2022.

[5] **A.L. Achu, C.D. Aju, Mariano Di Napoli, Pranav Prakash, Girish Gopinath, E. Shaji, Vinod Chandra**, “Machine-learning based landslide susceptibility modelling with emphasis on uncertainty analysis”, Geoscience Frontiers, vol. 14, 2023.

[6] **B. Veeresha Gowda, Sindhu B N, Tanuja S, Tejashwini More**, “EARLY LANDSLIDE DETECTION USING IOT”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 13, 2024.

[7] **Ravi Bhushan Bhardwaj**, “Landslide Detection System: Based on IOT”, International Journal for Scientific Research and Development, vol. 9, 2021.

[8] **Nitin Kumar M and Maneesha Vinodini Ramesh**, “Accurate IoT Based Slope Instability Sensing System for Landslide Detection”, IEEE Sensors council, vol. 22, 2022.

[9] **Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu**, “LightGBM: a highly efficient gradient boosting decision tree”, International Conference on Neural Information Processing Systems, 2017.

[10] **Scott M. Lundberg and Su-In Lee**, “A unified approach to interpreting model predictions”, International Conference on Neural Information Processing Systems, 2017.