Resource Recovery Optimization and Crisis Response Using Machine Learning and Big Data Tools

Annapoorna K K[1], Anila C Biju[2], Emlin Maria Roy[3], Prof. Ms. Jasmin T. Jose[4]
[1,2,3,4]*School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India*

Abstract—The Resource Recovery Optimization and Crisis Response Using Machine Learning and Big Data Tools project leverages advanced machine learning algorithms such as en-semble learning algorithms and big data analytics to enhance disaster and pandemic management. It focuses on improving disaster prediction. In pandemic scenarios, machine learning is utilized for predictive modeling, outbreak monitoring, and disease diagnosis. The project trains models using historical and real-time data to improve prediction accuracy and resource allocation efficiency. In addition, it addresses key challenges, including data reliability, infrastructure costs, and real-time processing constraints. By exploring synergies between machine learning and big data analytics, this initiative seeks to develop a centralized platform that supports both government and social efforts in effective crisis management and resource recovery.

Index Terms—Support Vector Machine, Random Forest Algorithm, Voting Classifier algorithm

I. INTRODUCTION

Landslides pose a significant threat to life, infrastructure, and the environment, particularly in regions prone to heavy rainfall, seismic activity, and unstable terrain. Traditional landslide prediction methods often rely on limited historical data and localized monitoring, making them less effective in providing timely and accurate risk assessments. With advances in big data analytics and machine learning, there is an op-portunity to enhance landslide prediction through real-time data processing and predictive modeling. This paper presents a novel approach that integrates multiple data sources, including sensor networks, satellite imagery, historical records, and real-time weather updates, to develop a comprehensive landslide prediction system. Using predictive analytics and machine learning algorithms, the proposed system analyzes key envi-ronmental factors such as rainfall patterns, soil moisture levels, slope of the terrain, and seismic activity to dynamically assess landslide risks. The ability to process vast data sets in real time enables early detection and timely alerts, allowing authorities to implement proactive measures for disaster preparedness and response. This approach not only improves the accuracy of landslide prediction but also enhances decision-making by facilitating efficient resource allocation and risk mitigation strategies.

II. LITERATURE SURVEY

The integration of machine learning and big data analytics in disaster and crisis response has emerged as a vital research area, driven by the necessity for real-time decision-making, predictive modeling, and optimized resource management. With the increasing frequency and intensity of natural dis-asters and global crises such as pandemics, the reliance on technology to provide scalable, accurate, and timely insights has become more critical than ever. Existing studies highlight various dimensions—from data acquisition and model training to infrastructure scalability and ethical concerns—providing a broad yet intricate view of how these technologies intersect to support crisis management.

[4] emphasized the utility of big data in disaster risk re-duction, leveraging real-time monitoring from IoT devices and insights from social media streams. These sources provide a continuous influx of data that, when processed effectively, can lead to timely alerts and reduced disaster impacts. Similarly, [2], [3] provided a comprehensive framework for incorpo-rating big data throughout the disaster lifecycle—including prediction, preparedness, response, and recovery—enabling governments and organizations to act with foresight rather than hindsight. offered a classification of data types such as satellite images,

call detail records (CDRs), and crowd-sourced reports, and matched them with processing frameworks like Hadoop and Apache Spark. [5] detailed methodologies to ensure data cleaning, quality assurance, and real-time responsiveness, which are essential when dealing with life-and-death situations.

[5] Machine learning models, such as Random Forest, Support Vector Machines (SVM), and deep learning architectures like Convolutional Neural Networks (CNNs), are widely adopted for disaster prediction and response. [6] used ConvLSTM models to enhance flood prediction accuracy, showing the importance of time series data in sequential modeling. [7] in-troduced a federated learning framework to decentralize model training and protect sensitive data, particularly beneficial in collaborative, cross-border crisis scenarios.

[6] Voting classifiers and ensemble methods have been exten-sively used to increase predictive performance by leveraging the strengths of multiple models. [8] documented their supe-riority in generalization and reliability, making them a prime candidate for resource-critical predictions.

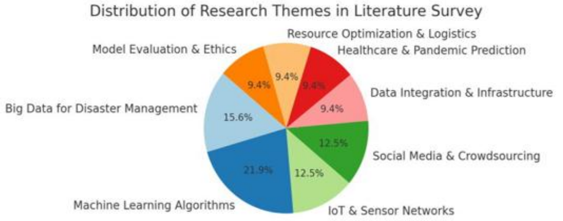

Fig. 1. Research themes in the survey

IoT devices are essential for real-time environmental data collection. Sensors measuring rainfall, temperature, soil mois-ture, and vibrations feed predictive systems with critical inputs.

[9] emphasized that combining official data with community-sourced sensor inputs improves accuracy and local relevance.

[10] examined landslide early warning systems powered by IoT, highlighting improved detection capabilities and reduced human risk.

Despite their potential, big data systems face significant hurdles in disaster applications. Data heterogeneity, noise, missing values, and privacy concerns complicate real-time analytics. [11] described these issues in urban data contexts, while [12] discussed the statistical techniques required to pre-process unreliable data for valid outputs.

Several successful implementations exist that demonstrate the practical viability of ML and big data in crisis contexts.

[13] described a real-time cyclone tracking system that uses distributed sensors and ML models for early warning. [14] predicted infrastructure damage due to seismic activity using geospatial features and building metadata. [15] focused on wildfire evacuations, showing how real-time GPS and mobile sensor data can inform safe routing strategies.

Big data frameworks such as Hadoop, Spark, and Apache Storm are foundational to handling high-velocity, high-volume data streams in crisis scenarios. [16] praised Spark's in-memory architecture for its real-time processing capabilities.

[17] outlined the role of cloud-based platforms in reducing infrastructure costs while maintaining scalability and fault tolerance.

Social media offers real-time, human-centric insights during disasters. [18] applied NLP to Twitter feeds to identify incident reports, affected zones, and public sentiment. [19] analyzed ethical and legal challenges related to misinformation and data usage from social media during crises.

The increasing integration of big data raises ethical ques-tions. [20] warned of the potential for surveillance and privacy breaches in data-rich disaster systems. [21] advocated for transparency and fairness in AI-driven decision-making to avoid biased resource allocation.

Disaster risk modeling often involves probabilistic methods.

[22] used Bayesian networks to simulate risk propagation in landslide-prone areas. [23] integrated CNNs with meteorolog-ical datasets to anticipate multiple disaster types. [24] showed that integrating satellite imagery with ML improves spatial prediction accuracy.

Efficient resource deployment can drastically reduce the impact of crises. [25] developed a deep learning model to predict regional resource demands during pandemics. [26] formulated a data-driven logistics optimization strategy. [27] proposed real-time resource distribution for smart cities based on IoT signals.

Model evaluation is critical for performance trustworthi-ness. Metrics such as precision, recall, F1-score, and ROC-AUC are used extensively. [28] emphasized robust validation techniques, and [29] reviewed evaluation metric selection, especially under class imbalance.

Research is expanding toward interpretable and collabora-tive AI. [7] proposed federated learning to preserve privacy in joint model training. [30] laid groundwork for explainable AI in disaster contexts. [31] suggested blockchain-based secure logging to ensure data integrity in sensor networks.

III. PROPOSED SYSTEM

The proposed system for landslide prediction using big data integrates multiple data sources, including sensor data, satellite imagery, historical records, and real-time weather information, to create a comprehensive predictive model. By leveraging advanced machine learning algorithms and predictive ana-lytics, the system analyzes various environmental factors, such as rainfall, soil moisture, terrain slope, and seismic activity, to assess landslide risk in real-time. The system's ability to process vast amounts of data enables early detection and provides timely alerts, facilitating proactive measures for disaster preparedness and response. This approach not only im-proves the accuracy of landslide prediction but also enhances decision-making, helping authorities allocate resources and issue warnings more effectively to mitigate potential damages.

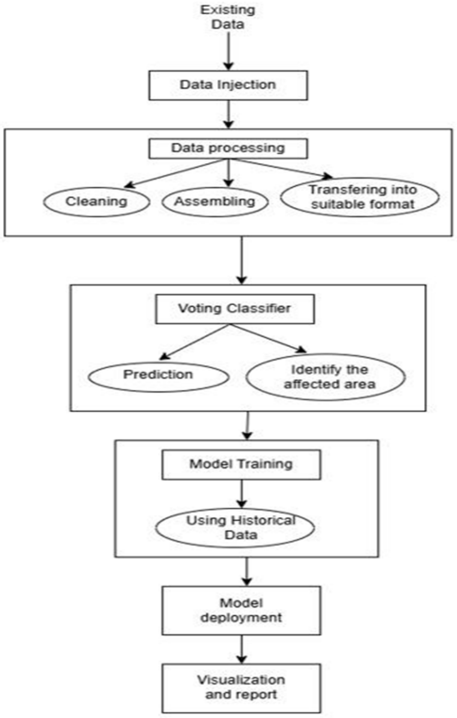

Fig. 2. Architecture diagram

A. Architecture

The architecture diagram depicts a comprehensive frame-work for landslide prediction using big data tools. The process starts with data injection from existing sources, including sensor data, geological records, and environmental datasets. This is followed by a data processing phase involving cleaning, assembling, and transforming data into a format suitable for analysis. A voting classifier is then applied to aggregate predic-tions from multiple models and identify affected areas prone to landslides. Historical data is utilized during the model training phase to improve the predictive accuracy. After training, the model is deployed to operational environments, where real-time prediction results are visualized and presented in reports for timely decision-making. The architecture leverages big data tools to handle large-scale, complex datasets, enhancing the reliability and efficiency of landslide prediction systems.

1) Data Injection: This stage integrates raw datasets from multiple sources, including historical disaster records, sensor data, and reports from authorities and volunteers. The collected data serves as the foundation for training machine learning models, ensuring comprehensive coverage of disaster-related variables. Proper data injection facilitates the seamless flow of information into subsequent processing stages, improving model performance and decision-making capabilities.

2) Data Processing: Data preprocessing is a crucial step that enhances the quality and reliability of the dataset. This phase involves:

Cleaning: Removing inaccuracies, duplicates, and irrelevant entries to ensure consistency.

Assembling: Merging multiple datasets into a unified format for streamlined analysis.

Transformation: Converting data into a machine-readable format compatible with machine learning models. Ensuring a well-processed dataset leads to more accurate predictions and mitigates errors caused by noisy data.

3) Voting Classifier for Prediction: A Voting Classifier is an ensemble technique that aggregates predictions from multiple models to improve performance. The models used include Logistic

Regression, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN).
Hard Voting: Predicts based on the majority class selected by individual classifiers.

4) Model Training: The models undergo training using his-torical disaster data, including past events, resource allocation patterns, and recovery timelines. Hyperparameter tuning is conducted using GridSearchCV to optimize parameters for each classifier:
Logistic Regression: Evaluates penalty (L1, L2) and regu-larization strength (C values).
KNN: Tunes the number of neighbors (k values).
SVM: Adjusts hyperparameters such as kernel type and regularization.
Random Forest: Optimizes the number of estimators for im-proved performance. This ensures that the best configurations are selected, maximizing predictive accuracy.

5) Model Deployment: Once trained and fine-tuned, the model is deployed into a production environment to facilitate real-time predictions. This deployment allows stakeholders to obtain timely insights into potential disaster-affected regions, assisting in proactive response and resource allocation.

6) Evaluation and Performance Metrics: The trained model is evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, a confusion matrix is plotted to analyze classification performance visually. The system's effectiveness is validated through rigorous testing on unseen data, ensuring generalization capability for real-world disaster scenarios.
This architecture emphasizes scalability, accuracy, and ac-tionable intelligence in resource allocation and disaster mit-igation. By leveraging machine learning, ensemble learning, and data-driven decision-making, the system provides robust predictions to aid disaster response teams and government agencies in mitigating the impact of disasters effectively.

IV. COMPARATIVE STUDY

The comparative analysis of various classifiers—Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and a Voting Classi-fier—highlights their performance differences based on pre-cision, recall, F1-score, and accuracy. Logistic Regression achieved the highest accuracy (0.9462) and excelled in pre-cision and recall for the 'yes' class, showcasing its ability to distinguish between classes effectively. Random Forest followed closely, with strong recall for the 'no' class (0.9692) and robust overall performance. KNN and SVM, while con-sistent, demonstrated slightly lower accuracy and F1-scores compared to the top performers. The Voting Classifier, an ensemble model, provided balanced results with an accuracy of 0.93, leveraging the strengths of its individual components. This makes the Voting Classifier a robust and generalizable option, especially for datasets with diverse patterns, even if it slightly lags behind Logistic Regression and Random Forest in accuracy.
The comparative analysis of various classifiers—Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and a Voting Classi-fier—provides a comprehensive insight into their strengths and limitations across precision, recall, F1-score, and accuracy metrics.
Logistic Regression emerged as the top-performing model with the highest accuracy of 94.62 percent, demonstrating exceptional precision (0.9833) and recall (0.9077) for the 'yes' class. This highlights its strength in effectively distin-guishing between classes and handling imbalanced datasets. It is particularly well-suited for applications requiring high interpretability and consistent performance across both classes.
Random Forest, with an accuracy of 93.85 percent, closely followed Logistic Regression. It demonstrated excellent recall for the 'no' class (0.9692) and robust precision, making it a reliable choice for scenarios requiring high recall to minimize false negatives. Its ensemble nature and ability to handle complex, non-linear relationships contributed to its strong performance.

Classifier	Precision (No/Yes)	Recall (No/Yes)	F1-Score (No/Yes)	Accuracy	Remarks
Logistic Regression	0.9143 / 0.9833	0.9846 / 0.9077	0.9481 / 0.9440	94.62%	Achieved the highest accuracy; excellent performance in precision and recall.
K-Nearest Neighbors (KNN)	0.9104 / 0.9365	0.9385 / 0.9077	0.9242 / 0.9219	92.31%	Balanced performance but slightly lower accuracy and recall for the 'yes' class.
Support Vector Machines (SVM)	0.8824 / 0.9194	0.9231 / 0.8769	0.9023 / 0.8976	90.00%	Consistent but underperformed in recall for the 'yes' class, lowering F1-scores.
Random Forest	0.9130 / 0.9672	0.9692 / 0.9077	0.9403 / 0.9365	93.85%	Strong recall for 'no' class; near-optimal F1-scores, highly reliable.
Voting Classifier	0.8900 / 0.9800	0.9800 / 0.8800	0.9300 / 0.9300	93.08%	Balanced precision and recall; robust generalization across datasets.

TABLE I
COMPARISON OF CLASSIFIER PERFORMANCE

K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) showed consistent but comparatively lower perfor-mance. KNN achieved an accuracy of 92.31 percent, with a balanced F1-score (0.9242 for ’no’ and 0.9219 for ’yes’). However, its reliance on the parameter n neighbors and sen-sitivity to data scaling slightly affected its results. SVM, with an accuracy of 90.00 percent, performed well in precision (0.8824 for ’no’ and 0.9194 for ’yes’), but its recall for the ’yes’ class was slightly weaker (0.8769), leading to lower F1-scores. This reflects the SVM’s sensitivity to hyperparameters like the regularization constant and kernel choice.

Finally, the Voting Classifier, an ensemble model combining the strengths of individual classifiers, achieved an accuracy of 93.08 percent. It maintained balanced precision (0.8900 for ’no’ and 0.9800 for ’yes’) and recall (0.9800 for ’no’ and 0.8800 for ’yes’), resulting in an F1-score of 0.9300 for both classes. While its accuracy was slightly lower than Logistic Regression and Random Forest, the Voting Classifier excelled in generalization and adaptability across diverse datasets. By integrating the decision boundaries of multiple classifiers, it minimized overfitting and performed robustly under varying patterns.

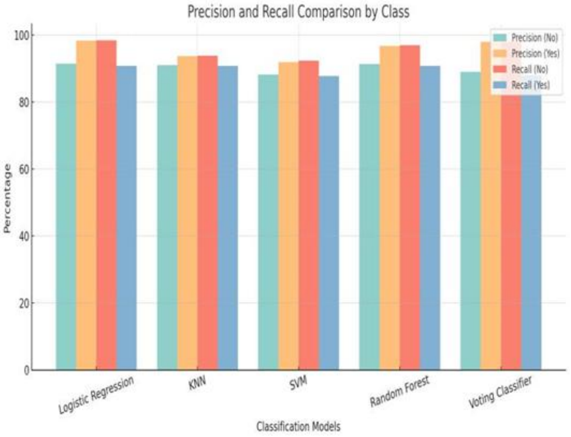

Fig. 3. ML comparison diagram

Overall, while Logistic Regression and Random Forest demonstrated higher accuracy, the Voting Classifier stood out as a robust and reliable choice for datasets with heterogeneous characteristics, striking a balance between precision and recall. Its ability to generalize well across multiple classes makes it a valuable tool in scenarios demanding consistent and reliable predictions.

V. CONCLUSION

The project highlights the power of Big Data analytics and machine learning in enhancing disaster preparedness and response. By multiple data sources and utilizing advanced analytical frameworks, the system efficiently processes real-time data to predict and allocate resources where they are needed most. This data-driven approach ensures optimized decision-making, reducing inefficiencies in resource distribution while improving response times during crises. The use of scalable Big Data frameworks allows the system to handle large volumes of information, making it adaptable to different disaster scenarios. While the project significantly enhances coordination between volunteers and government authorities, future improvements could focus on expanding data sources, refining predictive models, and improving system scalability to further strengthen disaster management efforts. Overall, the project serves as a crucial step toward leveraging technology for effective resource recovery and crisis response.

ACKNOWLEDGMENT

Every work is successful due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. At this juncture, we feel deeply honored in expressing our sincere thanks to Head of the Department Mr. Murali S, our project guide Mrs. Jasmine T Jose, Professor for the guidance, monitoring, and constant encouragement throughout the course. We also like to thank all the faculty members of the Computer Science and Engineering Department for their advice and guidance for the project. Lastly, we thank the almighty, our parents and friends for their constant encouragement without which this work would not be possible.

REFERENCE

[1] Yu, K. and Yang, C. (2018). Big data in natural disaster management: a review. Geosciences, 8(5), 165.

[2] Akter, S. and Wamba, S.F. (2019). Big data and disaster management: a systematic review and agenda for future research. Annals of Operations Research, 283(1), 939–959.

[3] Akter, S. and Wamba, S.F. (2019). Big data analytics in disaster response and recovery: A systematic review. Annals of Operations Research, 283(1–2), 939–959.

[4] Khalid, M.A., Roxin, A., Cruz, C., and Ginhac, D. (2017). A review on applications of big data for disaster management. In 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 370–377.

[5] Song, X. et al. (2022). Big Data and Emergency Management: Concepts, Methodologies, and Applications. IEEE Transactions on Big Data, 8(2), 397–419.

[6] Moishin, M. et al. (2021). Designing Deep-Based Learning Flood Forecast Model with ConvLSTM Hybrid Algorithm. IEEE Access, 9, 50982–50993.

[7] Farooq, M.S. et al. (2023). FFM: Flood Forecasting Model Using Federated Learning. IEEE Access, 11, 24472–24483.

[8] Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33(1), 1–39.

[9] Schempp, T. et al. (2018). An Integrated Crowdsourced Framework for Disaster Relief Distribution. In 5th Int. Conf. on ICT for Disaster Management, pp. 1–4.

[10] Zhou, X. et al. (2019). An IoT-enabled landslide early warning system based on LoRa and machine learning. Sensors, 19(6), 1236.

[11] Grolinger, K. et al. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. Journal of Cloud Computing, 2(1), 1–24.

[12] Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction. In Big Data Analytics, pp. 121–150.

[13] Shah, A. et al. (2019). Machine learning for cyclone intensity estimation using satellite imagery. Natural Hazards, 98(1), 27–49.

[14] Lin, G. et al. (2020). Earthquake damage prediction using geospatial ML and remote sensing. Natural Hazards Review, 21(3), 04020022.

[15] Greenwood, F. et al. (2017). Evacuation decision support using real-time analytics. In ISCRAM.

[16] Zaharia, M. et al. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65.

[17] Chen, M. et al. (2014). Big Data: A Survey. Future Generation Computer Systems, 37, 289–299.

[18] Imran, M. et al. (2015). Social media data for disaster response: An overview. In Social Informatics, pp. 1–10.

[19] Poblet, M. et al. (2014). Disaster management and social media: Twitter use during the Haiti earthquake. Computer Law & Security Review, 30(6), 679–688.

[20] Tene, O. and Polonetsky, J. (2013). Big data for all: Privacy and user control. Stanford Law Review Online, 66, 63–69.

[21] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Phil. Trans. of the Royal Society A, 376(2128), 20180089.

[22] Torres, R. et al. (2018). Bayesian network modeling of landslide susceptibility. Landslides, 15(3), 453–467.

[23] Jia, H. et al. (2020). Multi-hazard risk prediction using CNNs. Remote Sensing, 12(8), 1266.

[24] Liu, W. et al. (2016). Remote sensing data fusion with ML for risk mapping. IEEE Trans. on Geoscience, 54(1), 134–143.

[25] Ahmad, M.W. et al. (2020). DL models for forecasting resource de-mands. Journal of Big Data, 7(1), 1–20.

[26] Li, C. et al. (2021). Logistics optimization using big data and analytics. Transportation Research Part C, 130, 103293.

[27] Rathore, M.M. et al. (2016). Real-time big data analytics for smart city apps. Future Generation Computer Systems, 63, 30–42.

[28] Boulesteix, A.-L. et al. (2013). A plea for neutral comparison studies in computational sciences. Briefings in Bioinformatics, 14(2), 132–142.

[29] Chicco, D. and Jurman, G. (2020). MCC over F1 score in binary evaluation. BMC Genomics, 21, 6.

[30] Arrieta, A.B. et al. (2020). Explainable AI: Concepts and challenges. Information Fusion, 58, 82–115.

[31] Wang, W. et al. (2020). Blockchain for data security in disaster man-agement. IEEE Internet of Things Journal, 7(5), 4310–4323.