Assessing Landslide Risk in the Kaduha-Gitwe Corridor

Abstract

Landslides present significant risks in Rwanda, particularly in the Western, Northern, and Southern Provinces, where steep terrain and shifting climatic conditions contribute to increased susceptibility. This study assesses landslide risk in the Gitwe-Kaduha Corridor, an area within the Nyanza District identified by the University of Rwanda as high risk due to low forest density and unstable slopes. Four predictive frameworks were implemented—Ordered Linear Model, Large Language Model, Neural Network Model, and Random Forest Model—among which the Random Forest Model demonstrated the highest validation and test accuracy rates at 52.8% and 51.6%, respectively. Findings from this study contribute to improving early warning systems and informing mitigation strategies for landslides. However, all models tended to underpredict high-risk cases, highlighting the need for further refinement through enhanced feature engineering, data augmentation, and alternative modeling approaches.

Introduction

Landslides are among the most severe natural disasters, often resulting in fatalities and irreversible damage, particularly in hilly regions worldwide (Nsengiyumva et al., 2018). In developing countries, landslides frequently impact localized areas, causing substantial loss of life and billions of dollars in property damage.

Rwanda's Western, Northern, and Southern Provinces experience high exposure to landslides, with the Gitwe-Kaduha Corridor identified by the University of Rwanda as a high-risk area due to low forest density and unstable slopes (Mind'je, Mules, & Tshimanga, 2019; Nsengiyumva et al., 2018). Climate projections from the Climate Monitoring International Partnership Phase 3 indicate increasing temperatures and more intense and frequent rainfall, exacerbating landslide susceptibility. Despite these risks, predictive research on landslide hazards in Rwanda remains limited (Nsengiyumva et al., 2018).

Study Objectives and Key Findings

A comprehensive understanding of landslide risk in Rwanda, particularly in the Gitwe-Kaduha Corridor, is essential for developing effective mitigation strategies. This study evaluates landslide susceptibility using seven key environmental factors identified from prior research and available data: soil class, soil depth, type of land coverage, area of land coverage, land coverage density, riverside proximity, and roadside proximity (Nsengiyumva et al., 2018; Mind'je et al., 2019).

Four predictive frameworks were implemented—Ordered Linear Model, Large Language Model, Neural Network Model, and Random Forest Model—to assess their effectiveness in predicting landslide risk. Among these, the Random Forest Model demonstrated the strongest performance. Despite these results, all models tended to underpredict, particularly in high-risk classifications. This poses a critical challenge, as it can lead to insufficient mitigation efforts, inadequate resource allocation, and increased landslide vulnerability. Addressing these limitations requires further research on enhanced feature engineering, data augmentation, and alternative modeling approaches to improve predictive accuracy.

Problem Statement

Assessing landslide risk in the Gitwe-Kaduha Corridor requires predictive models that can effectively analyze available geotopical factors to identify high-risk areas. However, selecting an appropriate modeling approach remains a challenge due to variations in data quality, terrain complexity, and model performance.

This study investigates the implementation of predictive frameworks based on the available data and evaluates their effectiveness in accurately assessing landslide susceptibility. Understanding the strengths and limitations of different models is critical for improving risk assessment and informing future research on landslide prediction in Rwanda.

Data Cleaning & Preparation

Before implementing landslide risk predictive frameworks, it was essential to merge and clean the spatial datasets to ensure consistency and accuracy. This section details the steps taken to integrate GeoPackage files, address spatial misalignment issues, and ultimately prepare the data for predictive modeling. Challenges involving the Digital Elevation Model (DEM) are also highlighted in this section.

GeoPackage Exploration

To develop landslide risk prediction frameworks, it was necessary to first merge and clean four relevant GeoPackage files containing data on elevation, soil, land coverage, and erosion control for the Gitwe-Kaduha Corridor.[1]

The primary challenge in merging the GeoPackage files was the misalignment of polygons across datasets, making integration difficult. To address this issue, a hexagonal grid with 500m edge lengths was generated to cover the entire corridor, and the centroid of each hexagon was

[1] soil_depth.gpkg, forest_coverage.gpkg, cover_type.gpkg, and erosion.gpkg

determined. For each centroid, the nearest polygon from each of the four GeoPackage files was identified. If the distance between the centroid and the identified polygon was less than 250 meters, the polygon's attributes were mapped to the centroid; otherwise, the attributes were marked as unavailable.

This approach produced a dataset of centroids, each containing column values from all four GeoPackage files. The 250-meter threshold, chosen as half the distance between neighboring centroids, ensured that each hexagon was populated with values taken from a reasonable proximity to its center. The hexagonal grid structure was selected for its ability to maintain uniform spacing between neighboring points, resulting in a more consistent spatial representation.

After merging the GeoPackage files, basic data cleaning was conducted to prepare the dataset for predictive modeling. Relevant predictor columns were selected based on prior research and data availability. Landslide risk was designated as the response variable, derived from the risk_cat column in the erosion GeoPackage.

The end result, used for predictive modeling, was a cleaned tabular dataset with eight columns:

- Seven predictor variables: soil class, soil depth, type of land coverage, area of land coverage, land coverage density, riverside, and roadside.
- One response variable: landslide risk.

DEM Exploration

A Digital Elevation Model (DEM) was also explored as an additional predictor for landslide risk assessment, with QGIS (Quantum Geographic Information System) utilized for its spatial analysis and mapping capabilities. The process involved importing a GeoTIFF (.tif) raster file of Rwanda into QGIS, followed by raster analysis to overlay slope data onto the elevation model. However, significant challenges arose in aligning the Coordinate Reference System (CRS) with the rest of the project, which operates under EPSG:4326 (WGS 84 - Latitude/Longitude). The initial CRS conversion resulted in zero values for certain areas, particularly within and around the designated corridor, creating inconsistencies in the dataset. This issue was critical as merging the DEM-derived slope data with other datasets introduced excessive distances between corresponding points, affecting the accuracy of spatial analysis. Two main CRS options were examined to resolve these discrepancies. The first was TM_Rwanda, the default CRS associated with the raster file, and the second was WGS 84 / UTM Zone 36S (EPSG:32736), a widely used reference system for mapping Rwanda. Ultimately, neither successfully aligned the extracted slope values with the latitude and longitude points in the other project files, preventing seamless integration of datasets. Further exploration of reprojection techniques and spatial interpolation methods may be required to accurately align DEM-derived slope data with different spatial datasets essential for landslide risk modeling.

Data Analysis

To assess landslide risk, a range of predictive modeling techniques was explored, including ordered linear models, a fine-tuned large language model (LLM), Neural Networks, and Random Forests. These models were selected to incorporate diverse methodological approaches and evaluated for their effectiveness in predicting landslides, and are presented below in order of lowest to highest accuracy rate. Unless otherwise specified, the dataset was divided into 60% training, 20% testing, and 20% validation to ensure a balanced evaluation of model performance.

Figure 1: Accuracy and Error Rates for Each Model

Model	**Validation Accuracy**	**Test Accuracy**	**Test Overprediction Rate**	**Test Underprediction Rate**
Ordered Linear Model	43.3%	41.4%	6.3%	52.2%
DistilBERT (LLM)	50.3%	50.7%	17.2%	32.2%
EDLT Neural Network	48.1%	51.0%	16.5%	32.5%
Random Forest Model	52.8%	51.6%	15.5%	33.2%

Ordered Linear Model

The Ordered Linear Model was selected to account for the ordinal nature of the target variable, risk category, which consists of multiple ordered classes. Standard logistic regression does not adequately handle ordinal data, so statsmodels' Ordered Model was used for classification. The Newton optimization algorithm was implemented in place of the more commonly used BFGS method due to its ability to achieve higher accuracy, despite greater computational expense (Lam, 2020).

A key challenge in this approach was the significant class imbalance, with the majority of observations concentrated in the Moderate risk category. To counteract this, SMOTE (Synthetic Minority Over-sampling Technique) was applied during preprocessing to balance the dataset by increasing the representation of underrepresented classes. This adjustment improved classification of minority categories but did not eliminate the model's tendency to misclassify areas as Moderate, likely due to its overrepresentation in the dataset. The confusion matrix (Appendix 1) further illustrates this issue, showing frequent misclassification into the Moderate category, most likely driven by its high prevalence in the data.

Despite the application of SMOTE, the Ordered Model exhibited a high underprediction rate, struggling to classify "High", "Very High", and "Extremely High" risk cases accurately. The final performance metrics (Figure 1) further emphasize these limitations, with underprediction remaining a persistent issue and overall accuracy reflecting the model's difficulty in capturing extreme risk levels. These results suggest that either insufficient data exists for higher-risk areas or that important features may be missing from the current library of available data, ultimately limiting the model's ability to make reliable predictions for the most vulnerable regions.

Large Language Model

To further explore alternative modeling approaches, a DistilBERT uncased model was fine-tuned and deployed for risk-category prediction. This transformer-based model was selected for its strong performance in sequence classification tasks, particularly those requiring sentence-level comprehension (Hugging Face, 2019). Additionally, DistilBERT offers high computational efficiency and fast processing speed, making it well-suited for large-scale classification tasks (Sanh et al., 2019). The uncased variant was used to simplify tokenization and reduce unnecessary complexity in model learning, as case sensitivity was not a critical factor for this application.

To prepare the input data for DistilBERT, each row in the cleaned tabular dataset was transformed into a single concatenated string representation of input variables. For example: "*landcover (dense forest), coverage_density (high (>70%)), class (6.0), area_class (area greater than 2ha), depth (>100cm), roadside (no), riverside (no)*". Each string was then tokenized using DistilBERT's tokenizer, and assigned a label corresponding to its risk category. The dataset was split according to industry standards: 80% training, 10% validation, and 10% testing. Fine-tuning was performed using the training and validation sets, while the test set was reserved for final evaluation.

Various weight decay values (1e–3, 5e–3, 1e–2, 2e–2) and learning rates (1e−5, 5e−5, 1e−4) were tested during fine-tuning to optimize model performance. The optimal configuration—a weight decay of 2e–2 and a learning rate of 5e−5—was selected based on highest test accuracy, minimal overfitting, and continued learning.

The validation and test accuracy rates (Figure 1) indicate moderate predictive performance, with their similarity suggesting strong generalization capability. However, the model's high underprediction rate, consistent with other models, remains a concern. As shown in Appendix 2, the model frequently misclassified "High" and "Very High" risk cases as "Moderate" or lower, reinforcing the underprediction issue. Moreover, it failed to correctly classify any instances of "Extremely High" risk, highlighting its difficulty in identifying the most severe cases. The implications of this high underprediction rate will be explored further in the conclusion.

Neural Network

The neural network analysis explored two distinct models: Spatial Regression Graph Convolutional Neural Networks (SRGCNN) and Convolutional Neural Networks for Categorical Data (EDLT). Between the two, EDLT outperformed SRGCNN in both validation and test accuracy. The SRGCNN model achieved a validation accuracy of 42.4%, but its test accuracy dropped to 40.8%, indicating potential overfitting. In contrast, EDLT demonstrated stronger generalization, with higher validation and test accuracy, as shown in Figure 1, further supporting its effectiveness in landslide risk prediction. These results indicate that EDLT was the more robust and reliable neural network model for landslide risk prediction.

The architectural differences between EDLT and SRGCNN help explain why EDLT outperformed SRGCNN in landslide risk prediction. The primary distinction lies in their feature transformation and learning processes. The SRGCNN model uses spatial weighting to adjust and refine feature values based on nearby nodes. This means that this model primarily gathers information directly from neighboring points, making it most effective when strong spatial relationships exist. In contrast, EDLT focuses on finding meaningful connections between features rather than spatial positioning. The model transforms data using correlation based techniques, rather than solely relying on directly neighboring points. Because EDLT captures global feature interactions rather than relying solely on local spatial structures, it is better suited for datasets where feature relationships are more important than spatial dependencies. Given that our dataset exhibited weak spatial structure, EDLT demonstrated stronger predictive performance than SRGCNN.

Despite achieving higher accuracy, EDLT exhibited a consistent tendency to underpredict risk levels, as shown in Appendix 3. This is evident in cases where "high risk" classifications were incorrectly labeled as "moderate risk." While the model showed relatively strong predictive performance, the high underprediction rate presents concerns, highlighting the need for further fine-tuning to improve risk classification, particularly for high-risk scenarios.

Random Forest Model

The Random Forest model was selected for its ability to handle nonlinear relationships and rank feature importance, making it well-suited for landslide risk assessment. Feature importance analysis revealed that riverside and roadside contributed minimally to predictive performance, and their removal led to a 0.2% improvement in accuracy. Hyperparameter tuning was performed using GridSearchCV to optimize performance.

Various techniques were also explored to address class imbalance, including class weighting, BalancedRandomForestClassifier, and Synthetic Minority Oversampling Technique (SMOTE). Still, these methods reduced accuracy to approximately 48%, suggesting they were ineffective for this dataset. Given the low accuracy, AUC was explored as a potential evaluation metric, particularly One-vs-One AUC, which measures the model's ability to distinguish between class pairs. A comparison between the initial model and those incorporating balancing techniques showed that AUC decreased after oversampling and reweighting, indicating that the original model, with optimized parameters, was more effective at distinguishing between classes despite the imbalance.

An analysis of classification errors revealed clear misclassification trends, as shown in Appendix 4. Moderate and High risk levels were frequently overpredicted, while "very high" and "extremely high" risk levels were often underpredicted, highlighting the model's difficulty in accurately distinguishing between risk categories. The final performance metrics, presented in Figure 1, further highlight these challenges, showing the model's tendency to misclassify higher-risk cases. While Random Forest demonstrated the strongest predictive accuracy among the evaluated models, these findings indicate that it still struggles to differentiate between risk levels, particularly for the most severe classifications.

Findings

Across all four models, test and validation accuracies ranged between 41.4% and 52.8%, with Random Forest achieving the highest performance, attaining a 52.8% validation accuracy and a 51.6% test accuracy.

As shown in Figure 1, all models consistently exhibited a bias toward underprediction rather than overprediction. This trend poses significant challenges, as underpredicting landslide risk could potentially lead to insufficient mitigation efforts, inadequate resource allocation, and increased vulnerability to landslide hazards. Given these implications, this study has identified that improving accuracy and reducing underprediction rates is a critical focus for future research.

Conclusion

Our study contributes to landslide risk assessment in Rwanda by evaluating multiple predictive models on key environmental and topographic factors. Despite moderate predictive accuracy, our analysis highlights the potential of some of these models for landslide prediction, given their opportunity to capture complex relationships in the data. The high rate of underprediction in all models serves as a critical limitation, highlighting the importance for further research into feature engineering, data augmentation, and alternative modeling techniques. Additionally, our work underscores the need for enhanced data collection, particularly regarding precipitation patterns, past landslide records, and slope and elevation, which are features identified in two prior research papers as being critical predictors (Mind'je et al., 2019; Nsengiyumva et al., 2018). Specifically, the lack of documentation for the provided dataset made it challenging to interpret the significance of certain features, potentially limiting their effectiveness in predicting landslides.

Next steps should entail refining and expanding the predictive modeling of landslide risk in Rwanda by addressing limitations and enhancing applicability. To achieve this, expanding the scope of analysis across Rwanda should be a priority. While this study focuses on the Gitwe-Kaduha Corridor, landslides pose a significant risk across all of Rwanda's hilly regions. Future work should extend this analysis to additional districts identified as high risk. Moreover, a critical step toward practical application is making our findings further accessible. This can be achieved through developing an interactive dashboard that visualizes landslide risk predictions, key contributing factors, and potential mitigation strategies. Such a tool can assist local governments, environmental agencies, and disaster response teams in resource allocation and risk assessment. Finally, given the challenges of underprediction and room for improvement on predictive abilities, exploring alternative modeling approaches is essential going forward.

Expanding the research scope and addressing these limitations will enhance the accuracy and practical value of landslide risk assessments in Rwanda. Ultimately, integrating improved models with better data and practical tools can support more effective mitigation strategies, helping to safeguard vulnerable communities and infrastructure from future landslides.

References

Baseline study and development of indicators and targets ... Rema. (2021, February). https://www.rema.gov.rw/fileadmin/user_upload/FLR_Mayaga_Project_Baseline.pdf

Hugging Face. (2019). DistilBERT: a distilled version of BERT. Retrieved from https://huggingface.co/docs/transformers/model_doc/distilbert

Lam, A. (2020, November 26). *BFGS in a Nutshell: An Introduction to Quasi-Newton Methods | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/bfgs-in-a-nutshell-an-introduction-to-quasi-newton-methods-21b0e13ee504/

Mind'je, R., Mules, M., & Tshimanga, P. (2019). Landslide susceptibility and influencing factors analysis in Rwanda. Environment, Development and Sustainability, 22(8), 7985–8012. https://doi.org/10.1007/s10668-019-00557-4

Mohan, A., Singh, D., & Srinivasan, K. (2020). Review on remote sensing methods for landslide detection using machine and deep learning. Transactions on Emerging Telecommunications Technologies, 32(7). https://doi.org/10.1002/ett.3998

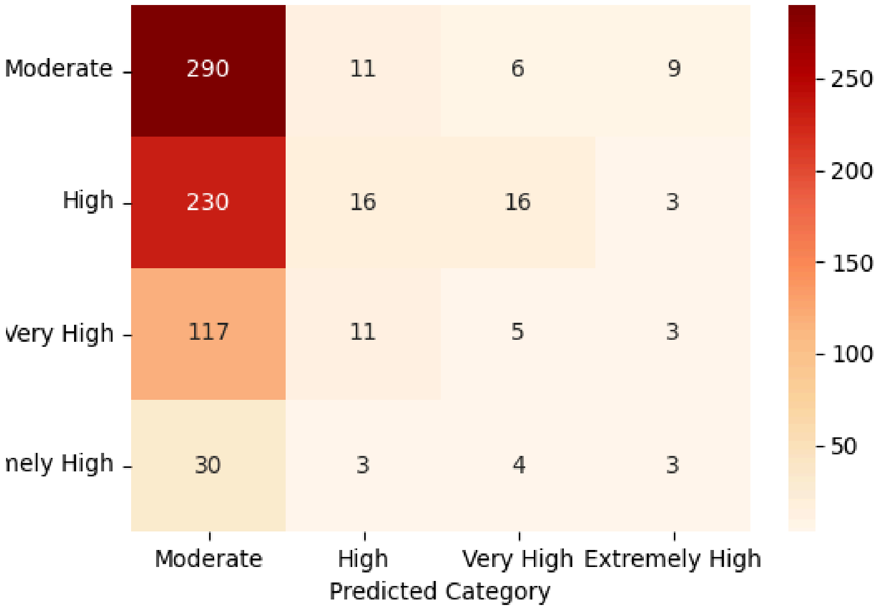
Nsengiyumva, J., Luo, G., Nahayo, L., Huang, X., & Cai, P. (2018). Landslide Susceptibility Assessment Using Spatial Multi-Criteria Evaluation Model in Rwanda. International Journal of Environmental Research and Public Health, 15(2), 243. https://doi.org/10.3390/ijerph15020243

Ordinal Regression — Statsmodels. (2023, December 14). Www.statsmodels.org. https://www.statsmodels.org/stable/examples/notebooks/generated/ordinal_regression.html

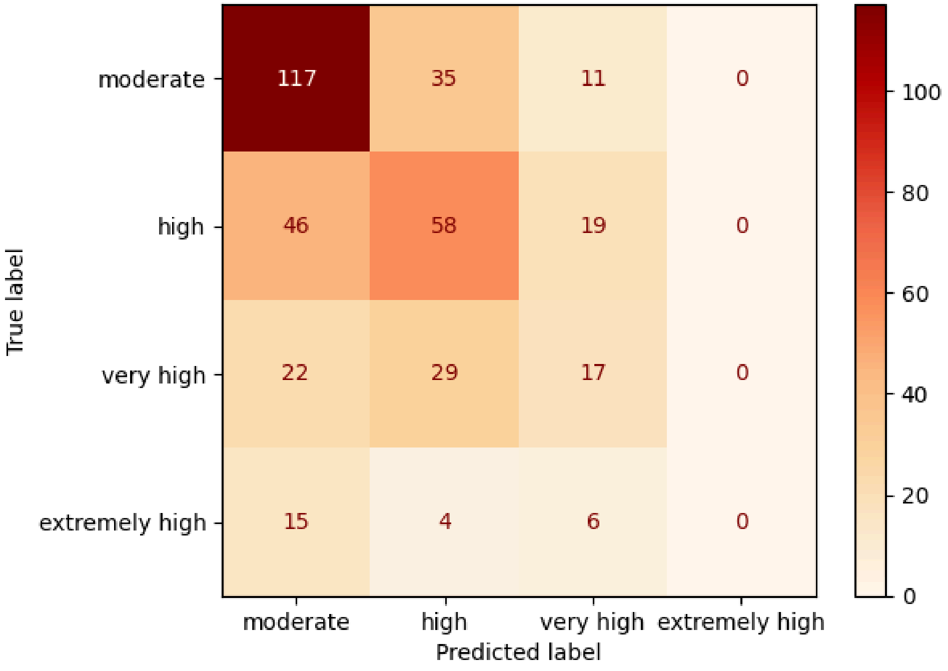
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Appendix

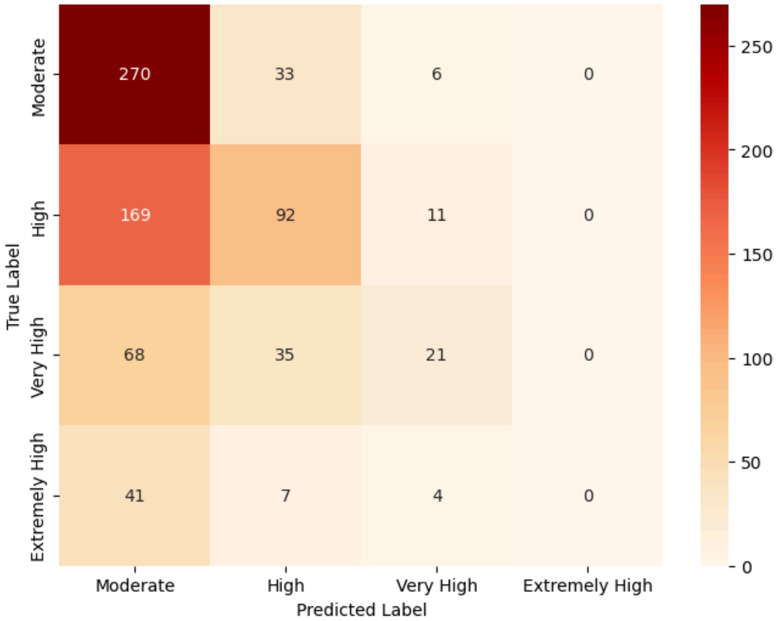
Appendix 1: Confusion Matrix for Ordered Linear Model (n = 758)



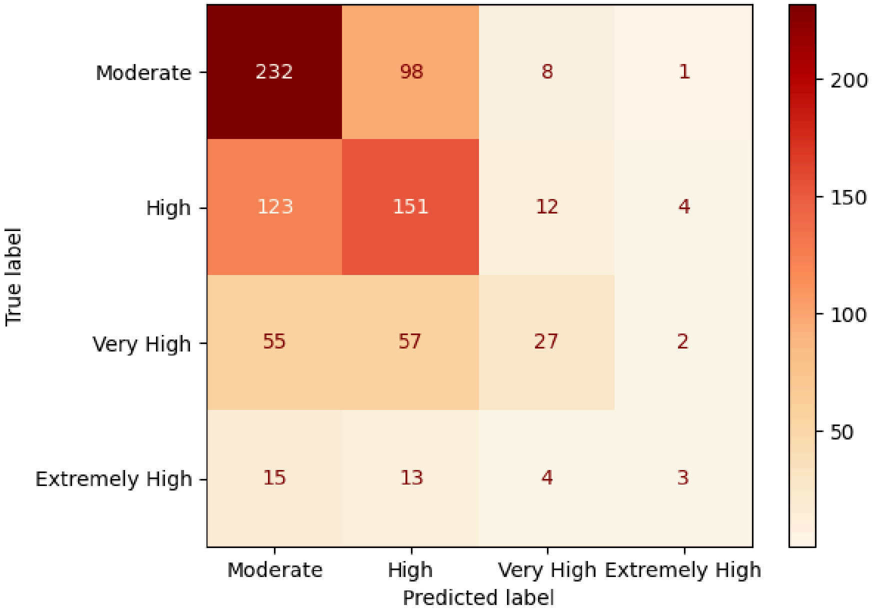
Appendix 2: Confusion Matrix (n = 379)



Appendix 3: Confusion Matrix for EDLT Model (n = 758)



Appendix 4: Confusion Matrix for Random Forest Model (n = 758)


True label
Moderate
High
Very High
Extremely High
Moderate
High
Very High
Extremely High
Predicted label