Open Access Original Article

Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis

Review began 11/07/2024
Review ended 12/02/2024
Published 12/04/2024

DOI: https://doi.org/10.7759/s44389-024-02268-y

Kaliprasanna Swain [1, 2] , Tan Kuan Tak [1] , Kamal Upreti [3] , Pravin R. Kshirsagar [4] , Sivaneasan Bala Krishnan [5] , Ramesh Chandra Poonia [3] , Sumya Ranjan Nayak [6] , Mihir Narayan Mohanty [7]

1. Electrical and Electronics Engineering, Singapore Institute of Technology, Singapore, SGP 2. Electronics and Telecommunication Engineering, Trident Academy of Technology, Bhubaneswar, IND 3. Computer Science, CHRIST University, Delhi NCR, Ghaziabad, IND 4. Computer Science, JD College of Engineering & Management, Nagpur, IND 5. Electrical Engineering, Singapore Institute of Technology, Singapore, SGP 6. Computer Engineering, KIIT University, Bhubaneswar, IND 7. Electronics, ITER, SoA University, Bhubaneswar, IND

Corresponding author: Kaliprasanna Swain, kaleep.swain@gmail.com

Abstract

Introduction: Accurately classifying stroke cases is a significant challenge in health care, as early detection can reduce severe complications and improve outcomes. Stroke datasets are usually imbalanced, with non-stroke cases in the majority, which poses a challenge to traditional machine learning algorithms and usually results in low stroke detection rates. This research proposes an advanced approach using Light Gradient-Boosting Machine (LightGBM) with Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors (SMOTE-ENN) to address this imbalance. Further optimization was performed using RandomizedSearchCV with LightGBM, achieving an area under the curve (AUC) of 0.946, higher than any of the baselines.

Methods: Class imbalance needed to be addressed first using SMOTE-ENN, which combines SMOTE with ENN for the creation of a balanced training set to be generalized well. We then used the LightGBM algorithm, which works quite efficiently with large datasets, and optimized it with RandomizedSearchCV.

Results: The highest performance of LightGBM was achieved with an AUC of 0.946, improving precision, recall, F1 score, and receiver operating characteristic curve performance. Therefore, LightGBM demonstrates high sensitivity and specificity for stroke detection compared to baseline models.

Conclusion: The integration of SMOTE-ENN with LightGBM and extensive hyperparameter optimization provides a robust framework for predicting stroke in imbalanced datasets. This approach not only enhances model performance but also proves to be a potential solution for other medical prediction problems significantly affected by class imbalance.

Categories: Image Processing and Analysis, IoT Applications, Data Science Methodologies
Keywords: stroke prediction, imbalance data set, smote-enn, lightgbm, randomizedsearchcv

Introduction

Prediction of a stroke is one of the highly important tasks in healthcare, as early detection can reduce the burden of serious complications and significantly improve outcomes [1]. However, accurate prediction of strokes is often compromised by inherent challenges, such as class imbalance in datasets [2]. These datasets are biased towards the majority class, which poses an obstacle for traditional machine learning models in identifying potential stroke cases.

Prediction accuracy has become a non-negotiable aspect of medical diagnosis, where the correct prediction of life-threatening conditions, such as stroke, is of utmost importance. Advanced models like this can enable timely medical intervention, saving lives and reducing disabilities in the long term [3]. This research, therefore, falls within the context of improving the predictive performance of machine learning algorithms when faced with imbalanced data, which is a common occurrence in most medical datasets, given the lower prevalence of positive cases (stroke) compared to negative cases.

The most common problem in machine learning is class imbalance, where instances of one class are reasonably higher in count than those of the other [4]. This creates a situation where models are biased towards the majority class, leading to poor identification of the minority class, which is often critical for detection in medical applications. While many traditional techniques, including Synthetic Minority Over-sampling Technique (SMOTE), may have artificially balanced these imbalances in the dataset, most of these methods fail in complex situations such as stroke prediction [5,6]. This paper, therefore, proposes an enhanced integrated approach: balancing class imbalances using SMOTE with Edited Nearest Neighbors (ENN) for the Light Gradient-Boosting Machine (LightGBM) algorithm and performing thorough tuning of model parameters.

How to cite this article
Swain K, Tak T, Upreti K, et al. (December 04, 2024) Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis. Cureus J Comput Sci 1 : es44389-024-02268-y. DOI https://doi.org/10.7759/s44389-024-02268-y

The analysis of various machine learning models in class imbalance challenges suggests that achieving high precision and recall, particularly for the minority class, is difficult. Techniques such as SMOTE fail completely in this case, as it could be noted that many models perform at a low precision, recall, and F1 score, especially when compared with the null accuracy-random forests and gradient boosting among them. This points to a critical gap in the existing methodology, as existing solutions are barely able to provide balancing factors between different metrics in highly imbalanced datasets. This paper highlights the development and tuning of a robust machine learning model designed to efficiently address the class imbalance problem. Using advanced techniques such as SMOTE-ENN, the work aims to improve the model's performance in correctly identifying and classifying minority classes, enabling the model to balance precision, recall, and F1 score. In conference or journal submissions, the goal is to prove that such techniques make an actual difference in terms of the whole predictive power/reliability of a model in real-life applications each involving data imbalance.

Related work

Rapid advancements have been made in stroke prediction using machine learning, driven by the challenging issue of imbalanced data and the demand for accurate predictive models. The contributions are listed chronologically in Table *1*, considering studies from 2019 to 2024, based on references[7-16]. They discuss the algorithmic methodologies employed, issues in handling class imbalances, and various optimizations of machine learning models to improve predictive accuracy for stroke prediction. The contents of each entry capture the main messages from all included studies, highlighting different methodologies and their impact on stroke identification in healthcare practice.

Ref. No.	Authors	Year	Key Contributions	Limitations
7	Liu et al.	2019	Introduced a hybrid ML method employing random forest and a DNN for optimizing stroke prediction on an imbalanced dataset, Overall accuracy = 71.6%.	Limited generalizability due to reliance on specific hybrid models; lacks robust handling of noise in data.
8	Wu and Fang	2020	Examined ML models' performance on imbalanced data among older Chinese, using techniques like SMOTE to improve accuracy, Accuracy = 78%.	Focused only on a specific demographic (older Chinese); performance metrics beyond accuracy not detailed.
9	Tazin et al.	2021	Proposed robust learning approaches for stroke prediction, favoring random forest for its high performance.	Limited evaluation of other advanced resampling techniques; lack of hyperparameter optimization.
10	Butt et al.	2022	Applied SMOTE Upsampling and optimized feature selection for predicting heart failure, methodologically similar to stroke prediction, accuracy = 84.11%.	Focused on heart failure, not directly validated for stroke prediction; potential overfitting with SMOTE.
11	Santos et al.	2022	Used AIS and decision trees via genetic programming for stroke prediction, addressing class imbalance innovatively.	Complexity of genetic programming may hinder scalability; limited comparative analysis with other methods.
12	Biswas et al.	2022	Explored multiple classifiers for stroke prediction, focusing on ROS to correct data imbalances.	ROS may lead to over-representation of minority class; no in-depth exploration of ensemble methods.
13	Wang et al.	2023	Explored AutoML and RUS for improving stroke prediction models in imbalanced datasets.	AutoML's black-box nature limits interpretability; RUS can lead to loss of informative data in majority class.
14	Dahiya et al.	2023	Examined gradient boost methods (XGBoost, LightGBM, CatBoost) highlighting the role of hyperparameter tuning and feature importance.	Limited to boosting algorithms; no exploration of resampling techniques for class imbalance.
15	Ushasree et al.	2024	Demonstrated a stacking methodology using various classifiers to enhance stroke prediction accuracy.	Stacking models can be computationally expensive and prone to overfitting without careful tuning.
16	Merdas	2024	Utilized EMS (Elastic Net–MLP–SMOTE) model to increase the performance of the model.	EMS model's performance may not generalize well to larger, diverse datasets; high reliance on SMOTE's synthetic data.

TABLE 1: Summary of literature

DNN: Deep Neural Network; ML: Machine Learning; SMOTE: Synthetic Minority Over-sampling Technique; AIS: Artificial Immune Systems; ROS: Random Oversampling; AutoML: Automated Machine Learning; RUS: Random Undersampling; LightGBM: Light Gradient-Boosting Machine

The objective of our study is to extend the current methods that predict stroke risk, using unique machine learning approaches aimed at addressing class imbalances in medical datasets - a problem encountered across most such scenarios. Our study benefits from using SMOTE-ENN and LightGBM, respectively, to improve the sensitivity of stroke prediction models that may suffer under skewed data distribution in traditional methods. Unlike conventional approaches which could present poor performance caused by highly imbalanced instances among classes, our work conducts a sophisticated combination of SMOTE and ENN together with classifying algorithm-specific improvements for approaching balanced samples on both sides - positive side as well as negative side. This adequately enhances not only specificity but also simultaneously maintains or even increases sensitivity aspects toward predicting future stroke outcomes.

This combined approach of upsampling and downsampling not only helps balance the guided majority classes but also improves classification results for minority instances, as these smaller class examples are critical for early stroke prediction. The use of RandomizedSearchCV for exhaustive hyperparameter tuning further optimizes the model's performance and leads to a first-of-its-kind attempt at this problem. Positioning our work within the context of broader research reveals that, while similar solutions exist, none tackle class imbalance with the same precision using advanced resampling and model optimization as we have done here - establishing a versatile scaffold adaptable to any imbalanced medical prediction scenario.

Materials And Methods

Materials

HEALTH 10650 is the dataset used in this study, a healthcare dataset focused on stroke prediction. It comprises 5,110 instances with 12 mixed-type features. Initially released as a medical research project, its purpose is to determine what features affect the likelihood of an individual experiencing a stroke. The dataset is used in binary classification, where 1 indicates a stroke (identified by MRI) and 0 indicates no stroke. It includes demographic health attributes such as age, gender, hypertension, heart disease, BMI, and average glucose level, along with lifestyle factors like smoking status and work type. The proposed sequential diagram for this work is shown in Figure 1.

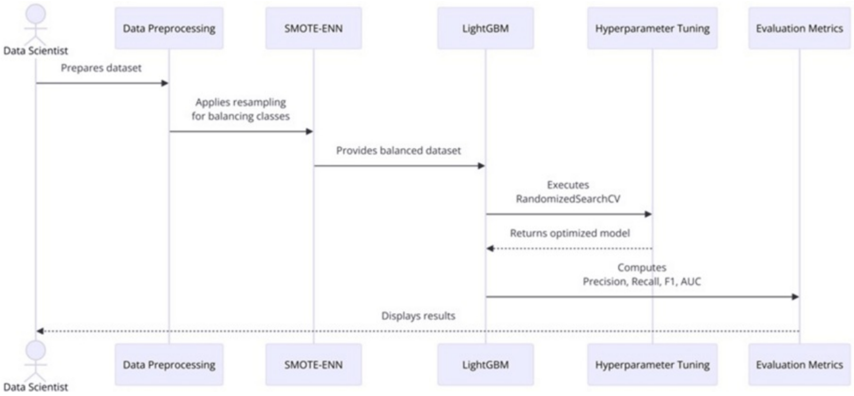

FIGURE 1: Sequence diagram for the proposed work

LightGBM: Light Gradient-Boosting Machine; SMOTE-ENN: Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors: AUC: Area Under the Curve

Preprocessing

Data preprocessing was performed before the analysis. Missing values, particularly in the BMI column, were handled through mean imputation to ensure completeness. Categorical variables such as gender, ever_married, work_type, residence_type, and smoking_status were converted into numerical format using one-hot encoding, making them suitable for machine learning algorithms. Similarly, numerical features were normalized using Min-Max scaling, setting the data within a fixed range so that all input variables were equally treated during training. The dataset was split into training and testing sets with an 80-20 split to measure model performance on new, unseen data.

Feature engineering

These techniques further enhanced stroke prediction with these models. In feature transformation and aggregation, polynomial features were generated for attributes such as age, avg_glucose_level, and BMI to uncover nonlinear relationships within the data. Outlier detection and removal were performed using interquartile range analysis, particularly for features like age and BMI that could contain extreme values, which might skew model predictions. Recursive feature elimination enabled dimensionality reduction, helping to select informative features for efficient models. Given the class imbalance in the stroke target variable, synthetic features were developed using SMOTE. This oversampled the minority class (patients who had a stroke), making the dataset more balanced and helping the model generalize better from these instances.

Experimental setup

The dataset was split such that 80% goes to the training set and 20% for testing, so that model testing could be done on unseen data. It can be observed that the models were trained on the training set, while in contrast, the testing set was used to assess the performance. In order to promote model generalization and avoid complex overfitting of the models, the training data was divided based on a five-fold cross-validation strategy. Performance metrics to be measured included accuracy, precision, recall, F1 score, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve to ensure that all analyses were well rounded. These metrics would provide further detail of model performance, particularly in a context where data is imbalanced, and examination based on accuracy may not be good enough.

Baseline algorithm performance on imbalanced data

First of all, the analysis was done to see how well conventional machine learning models perform on the imbalanced dataset. The models were chosen for this purpose because logistic regression, random forest, and gradient boosting are very common in classification. It was about setting up a baseline in performance that could then be improved upon using advanced techniques.

These results are disappointing, as can be seen in Figure2 and Table 2; this is particularly the case when considering how the imbalanced data was handled. The dataset was highly skewed, where the majority class occupied 95% of all instances, giving a null accuracy of 95%. That said, none of the models performed satisfactorily. In fact, the worst performance was that of the random forest model, which almost showed zero precision and recall for the minority class in its real and utter failure to predict instances from that class correctly. Indeed, all models had this result consistently, as told by the low F1 scores and AUC values.

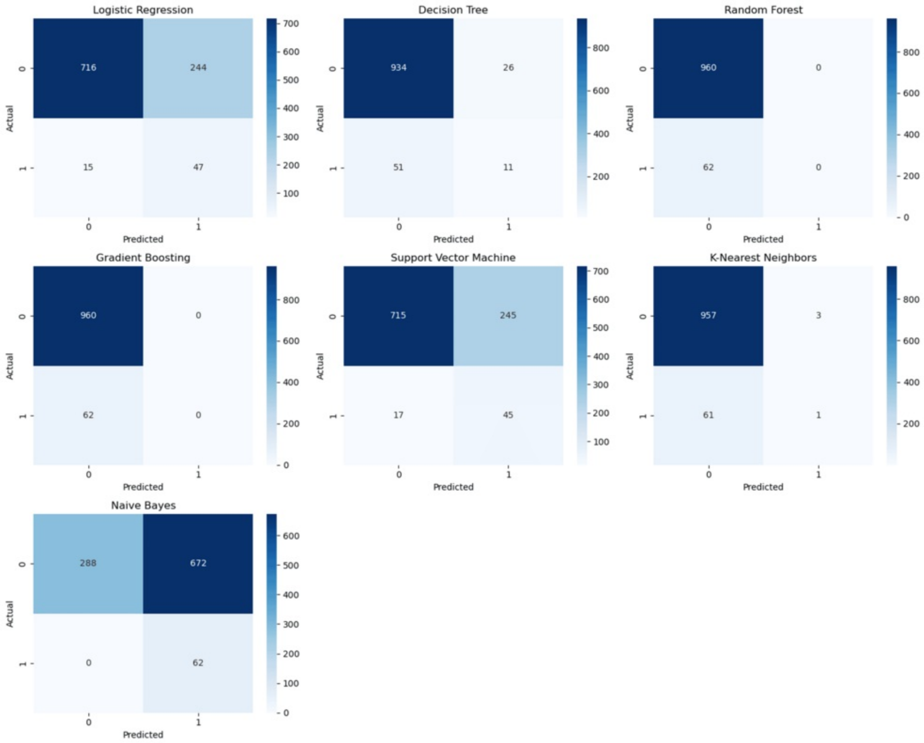

FIGURE 2: Confusion matrix for baseline algorithms

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.745597	0.160959	0.758065	0.265537	0.854032
Decision Tree	0.924658	0.317073	0.209677	0.252427	0.590255
Random Forest	0.939335	0	0	0	0.804024
Gradient Boosting	0.940313	1	0.016129	0.031746	0.827797
Support Vector Machine	0.74364	0.155172	0.725806	0.255682	0.817776
K-Nearest Neighbors	0.937378	0.25	0.016129	0.030303	0.645657
Naive Bayes	0.342466	0.084469	1	0.155779	0.835786

TABLE 2: Performance analysis of baseline classifiers

AUC: Area Under the Curve

Figure *3* compares all these models, indeed, showing that even though the logistic regression and gradient boosting performed well - training AUCs of 0.854032 and 0.827260, respectively - their overall performance was still abysmal. Their precision and recall were way below accepted values, especially for the minority class. That pointed to a crucial limitation of the baseline algorithms: they could tell the classes apart somehow - as reflected in the AUC scores - but totally failed to handle the class imbalance issue, since their predictive performance was far from good on the minority class. This clearly showed from the analysis that there was a need for the adoption of more sophisticated techniques that could surmount the adverse effects brought about by class imbalance; thus, the next phase of the study implemented and evaluated resampling techniques to improve model performance.

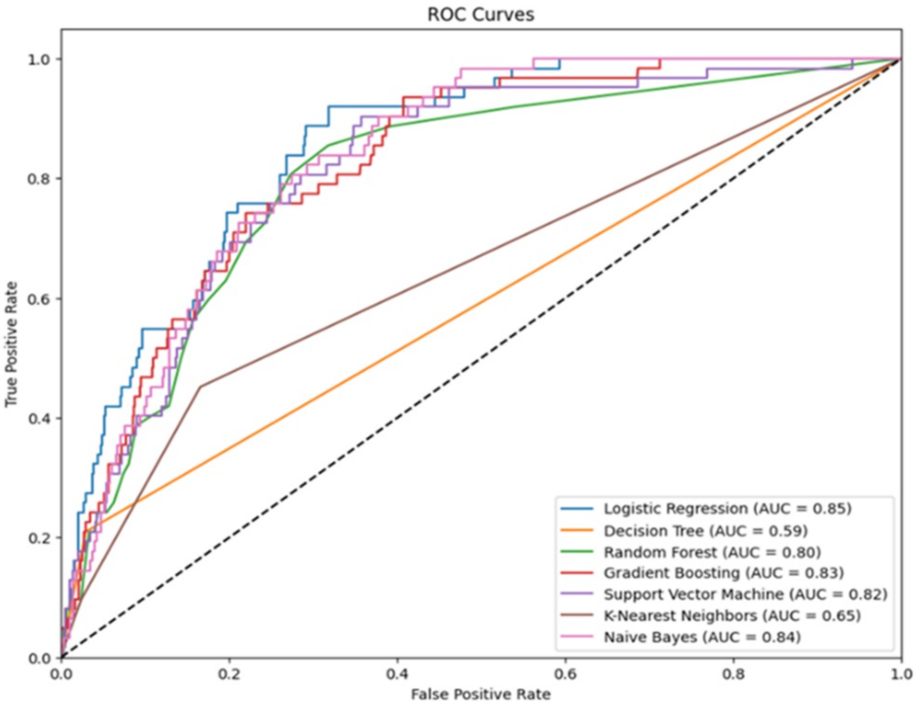

FIGURE 3: ROC curve for baseline algorithms

ROC: Receiver Operating Characteristic; AUC: Area Under the Curve

Application of SMOTE for dataset balancing

To handle this gap in the baseline models, we developed SMOTE. One of the well-accepted methods for generating synthetically prepared samples of the minority class to balance out the dataset is SMOTE. The idea behind using SMOTE is that increasing the representation of the minority class would allow the models to learn from a more balanced distribution of the data. The models were then evaluated again after over-sampling with SMOTE. As shown in Figures 4 and 5, after the use of SMOTE, there is a considerable improvement concerning the classification of the minority class. Logistic regression had notably better recall, which in fact means the model was much more sensitive to finding the instances of a minority class. Logistic regression AUC value increased, which means an improved capability to distinguish between the two classes. With the improvements, challenges still remained. Table 3 depicts an illustration where, even though the recall considerably improved, precision did not improve proportionately. That is to say, while the models improved at detecting the minority class, there was still a leaning toward false alarms. Especially, in the case of the gradient boosting model, while improving in AUC, it struggled to balance precision and recall.

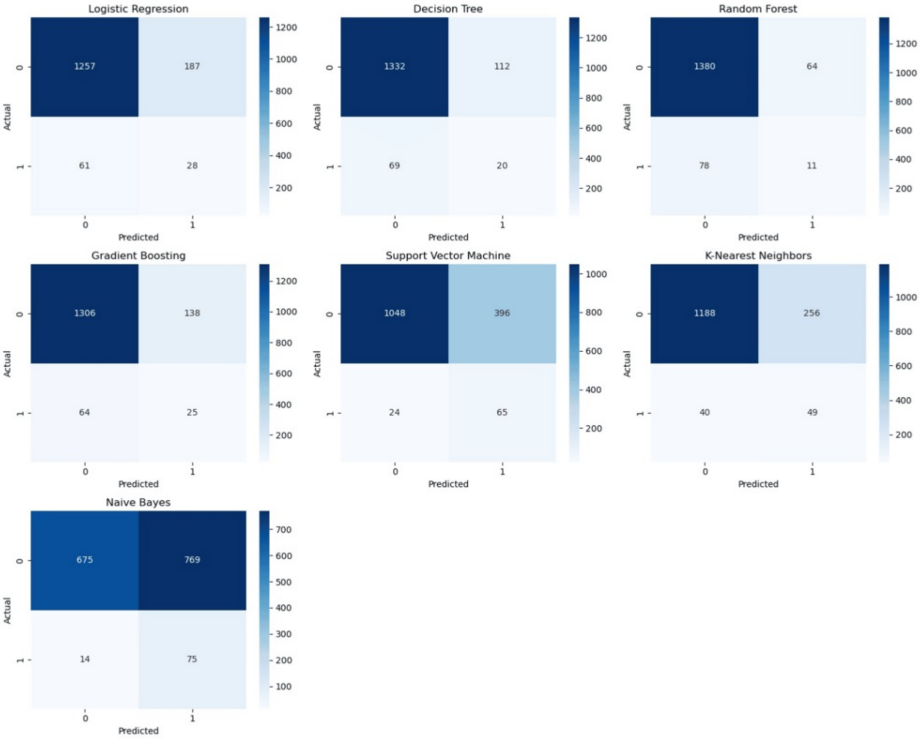

FIGURE 4: Confusion matrix for baseline algorithms after applying the SMOTE

SMOTE: Synthetic Minority Over-sampling Technique

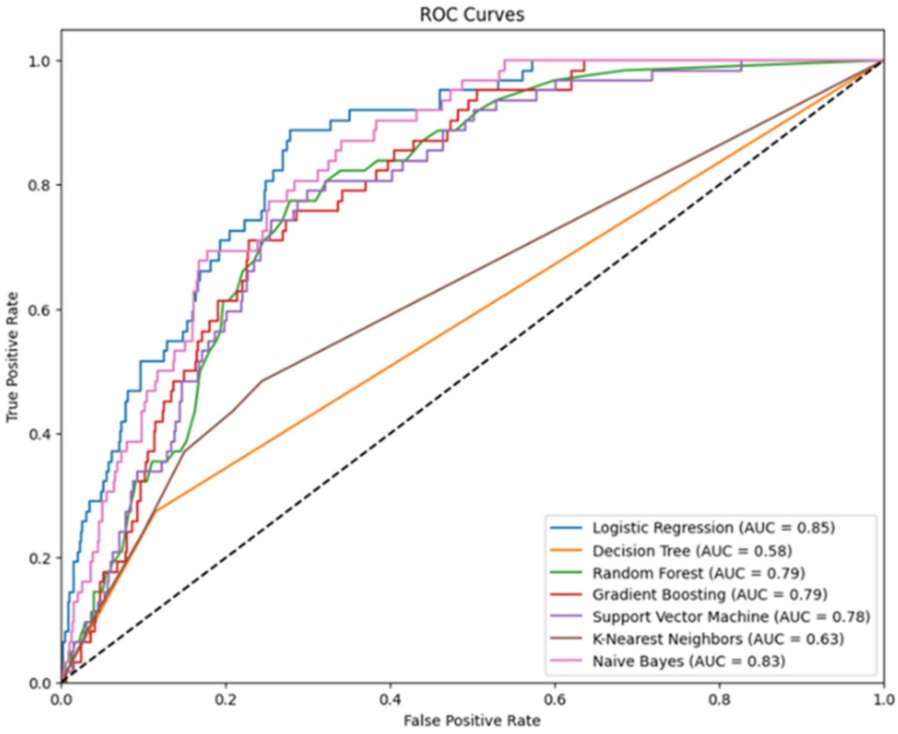

FIGURE 5: ROC for baseline algorithms after applying the SMOTE

ROC: Receiver Operating Characteristic; SMOTE: Synthetic Minority Over-sampling Technique; AUC: Area Under the Curve

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.740705	0.167213	0.822581	0.277929	0.851008
Decision Tree	0.848337	0.133858	0.274194	0.179894	0.579805
Random Forest	0.901174	0.157895	0.145161	0.151261	0.788248
Gradient Boosting	0.835616	0.180723	0.483871	0.263158	0.793145
Support Vector Machine	0.793542	0.159817	0.564516	0.24911	0.784157
K-Nearest Neighbors	0.820939	0.137725	0.370968	0.208873	0.628663
Naive Bayes	0.337573	0.083897	1	0.154806	0.83377

TABLE 3: Outcomes for baseline algorithms after applying the SMOTE

AUC: Area Under the Curve; SMOTE: Synthetic Minority Over-sampling Technique

Enhanced resampling with SMOTE-ENN

In order to extend performance and balance the precision and recall of the model, we utilized the SMOTE-ENN technique. SMOTE-ENN is an extension of SMOTE, combined with ENN. While SMOTE focuses on generating synthetic examples of the minority class, ENN removes noisy or ambiguous examples in both the minority and majority classes. The dual approach here in this case balances the dataset, cleans it, and hence reduces the chance of a model learning from misleading data points.

The results over Figure 6 indicate that the application of SMOTE-ENN had a profound impact on model performance. The LightGBM model, in particular, showed marked improvements across all metrics. The AUC increased to 0.9330, and there was a balanced enhancement in both precision and recall, which led to a significant improvement in the F1 score. This suggests that the model was not only better at identifying the minority class but also more accurate in doing so, reducing the rate of false positives. The fine-tuning of the LightGBM model further optimized its performance. By adjusting hyperparameters through grid search, we were able to achieve an AUC of 0.9457, as shown in Figure 7. This fine-tuning process also improved recall and F1 scores, particularly for the minority class, demonstrating that the model became more adept at handling the nuances of the imbalanced dataset.

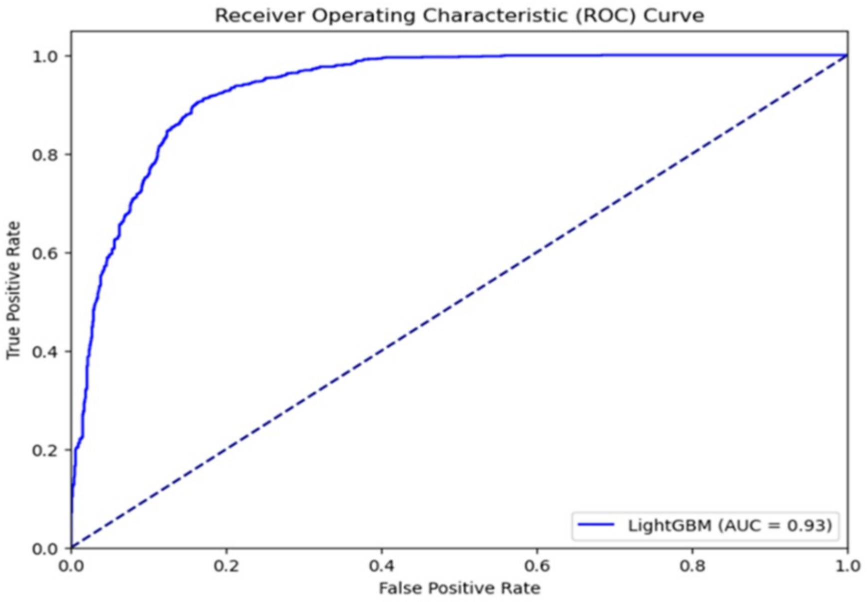

FIGURE 6: Receiver operating characteristic curve for LightGBM algorithm after applying SMOTE-ENN

LightGBM: Light Gradient-Boosting Machine; SMOTE-ENN: Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors: AUC: Area Under the Curve

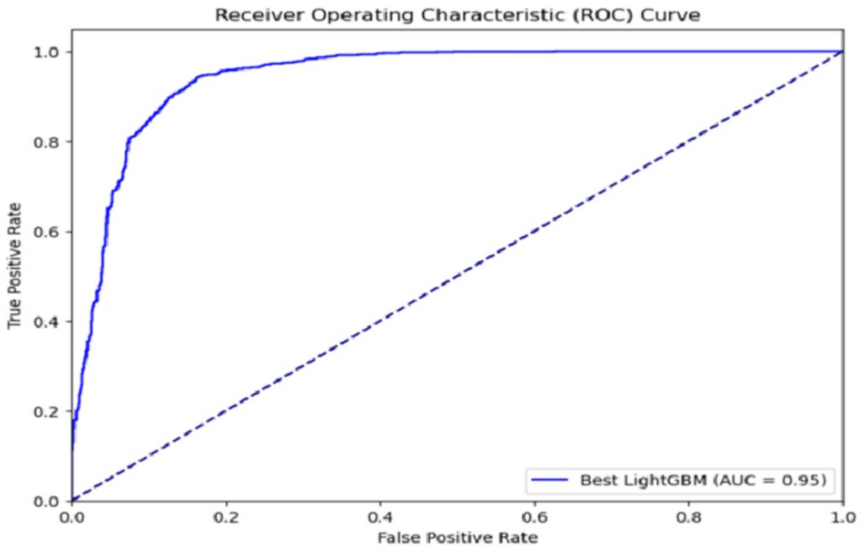

FIGURE 7: Receiver operating characteristic curve for LightGBM algorithm after fine tune

AUC: Area Under the Curve; LightGBM: Light Gradient-Boosting Machine

Visualizing model performance

The following models have been developed in order to compare the comprehensive performance of models, and then several performance curves were produced and analyzed:

Confusion Matrix: Figure *8* illustrates the detailed information on the model prediction and the numbers of true positives, true negatives, false positives, and false negatives. The balance between sensitivity and specificity is much better for the model LightGBM when using the SMOTE-ENN technique in the data preprocessing, since it reduces the number of false negatives as low as possible and, at the same time, it holds the reasonable rate of false positives low.

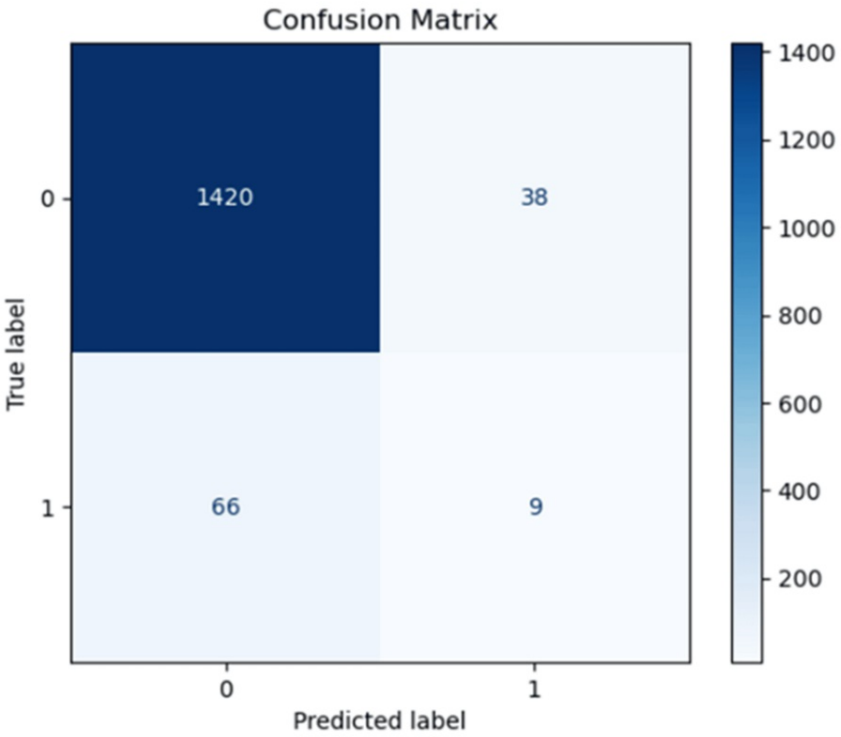

FIGURE 8: Confusion matrix of LightGBM

LightGBM: Light Gradient-Boosting Machine

Threshold vs. F1 Score: Figure *9* depicts that on changing the decision threshold, there is a variation in the F1 score. This will be important to be able to understand how the changes in thresholds will affect the balance between precision and recall. The optimal threshold at which LightGBM best shows its value of F1 score gives the best trade-off between precision and recall.

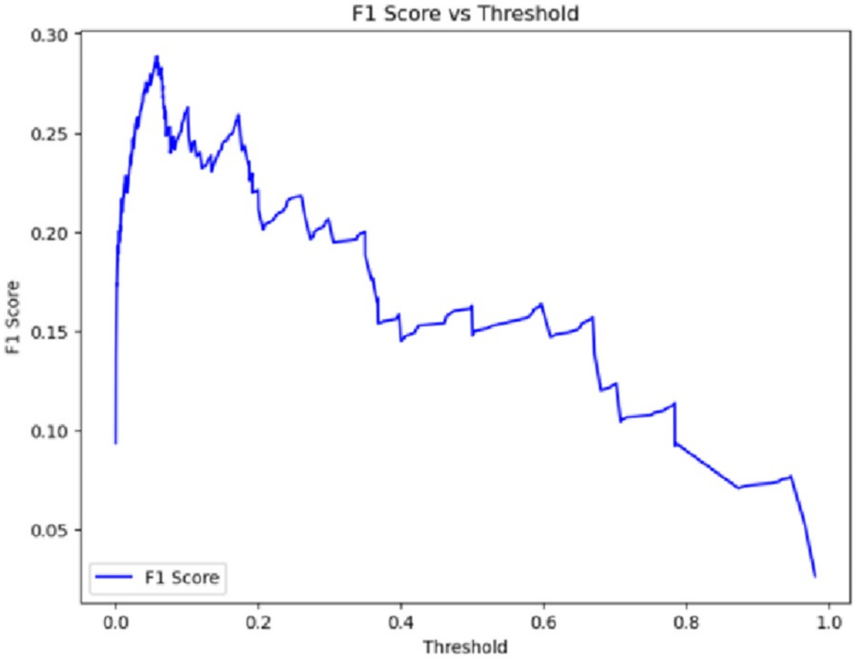

FIGURE 9: F1 score vs. threshold

Calibration Curve: Figure 10 shows the calibration curve that gives an idea about the match between the predicted probabilities and actual outcomes concerning the LightGBM model. A well-calibrated model should give a curve that is close to the diagonal, reflecting that the predicted probabilities are accurate representations of the true likelihood of an event. Application of SMOTE-ENN helped in aligning the predicted probabilities closer to this ideal, demonstrating improved calibration.

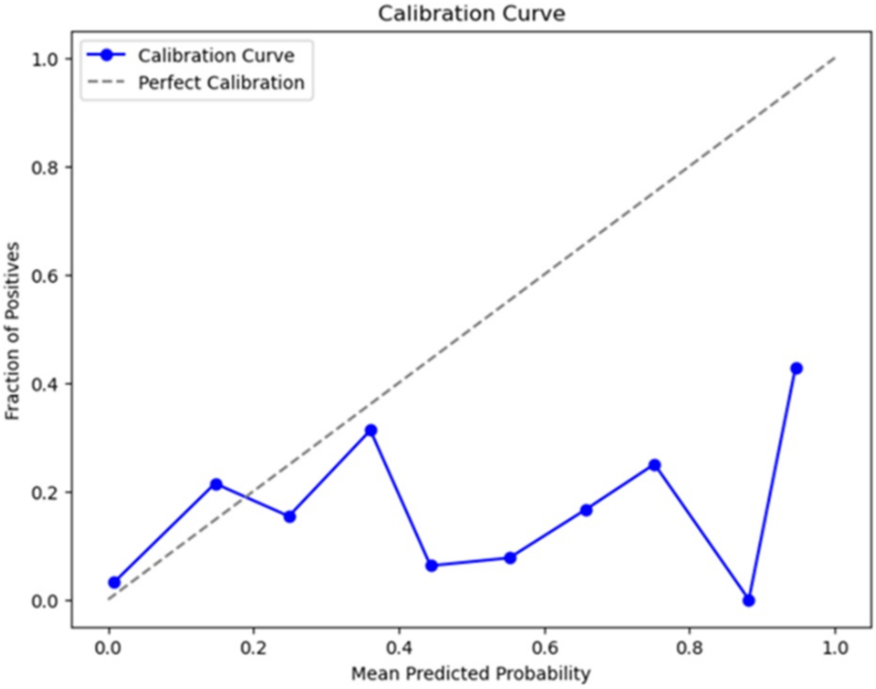

FIGURE 10: Calibration curve

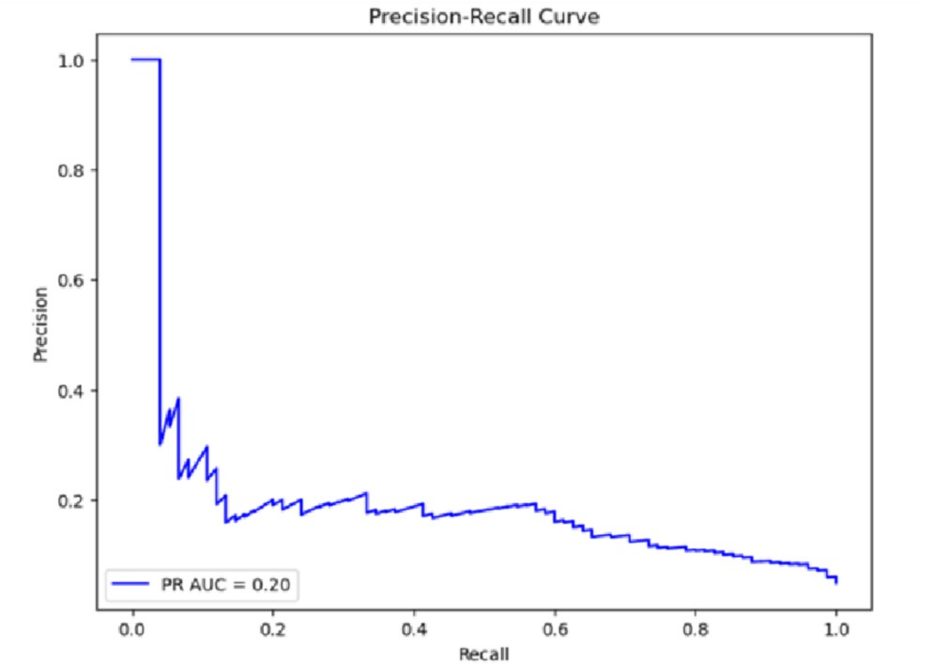

FIGURE 11: Precision-recall curve

PR: Precision-Recall; AUC: Area Under the Curve

Precision-Recall Curve: Figure *11* is a precision-recall curve that gives a good view of the trade-off between precision and recall. This curve is informative, especially in datasets where there is class imbalance, because precision and recall provide more useful insight compared to accuracy alone. In those cases, LightGBM was able to sustain a better balance between precision and recall after the application of SMOTE-ENN - a very essential thing for keeping the false positives as low as possible while capturing the majority of positive cases.

Cumulative Accuracy Profile (CAP) Curve: The CAP curve in Figure*12* emphasizes how well an improvement in true positives can be captured with fewer false positives. This is further solidified by how distinct the LightGBM model CAP curve outperforms those of the baseline models after applying the advanced resampling technique.

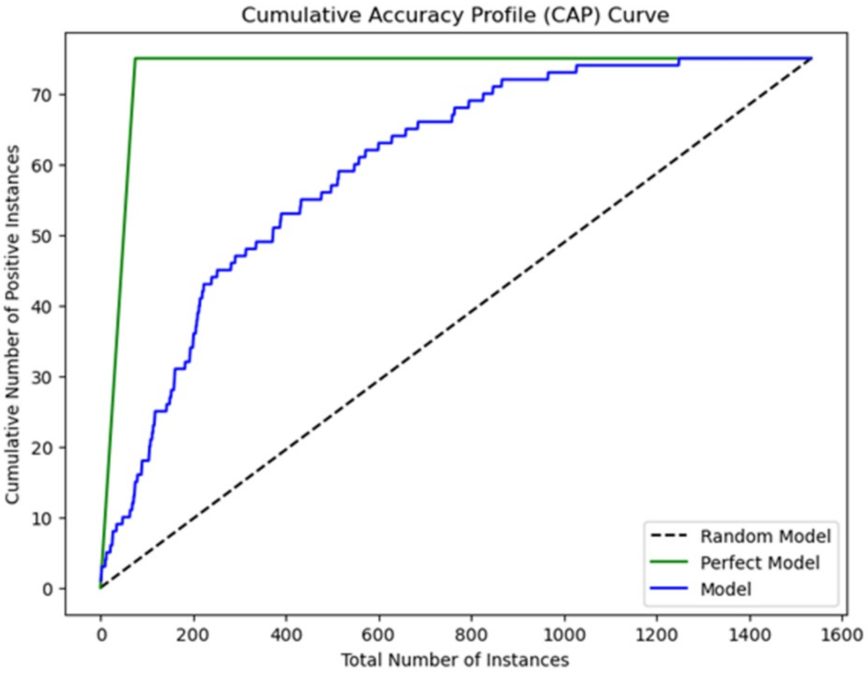

FIGURE 12: Cumulative accuracy profile curve

Learning Curve: Figure 13 presents the learning curve of the LightGBM model, which generally reflects how model performance increases with more training data. The learning curve illustrates that with the especially applied SMOTE-ENN, the LightGBM model kept learning well from the data without significant overfitting and, hence, showed good generalization on new data.

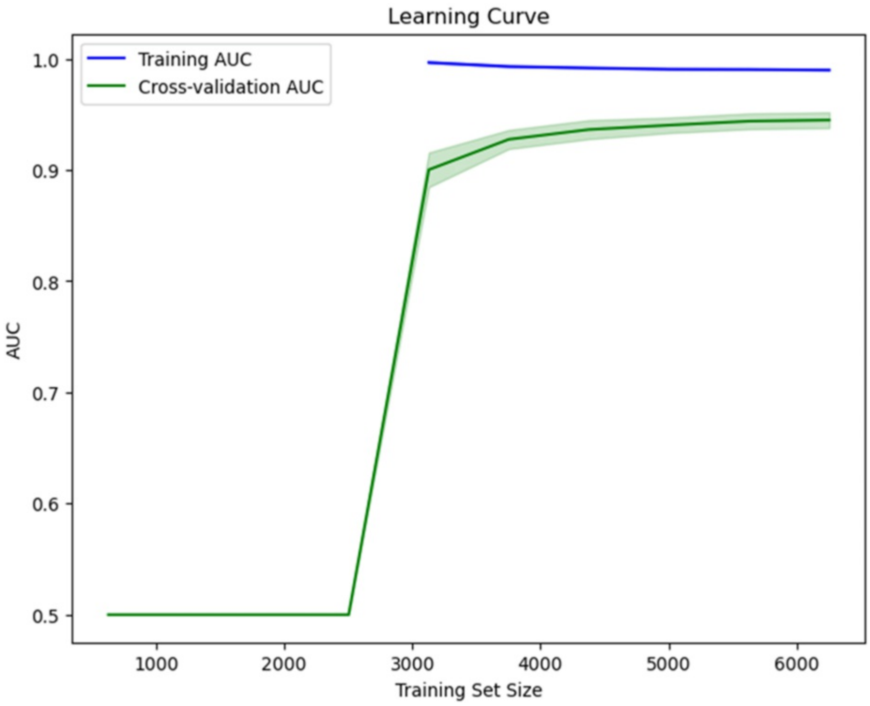

FIGURE 13: Learning curve

AUC: Area Under the Curve

Lift and Gain Chart: Figure 14 shows the gain and lift chart, which details how well the model is performing in predicting true positives. It is observed that LightGBM, after performing SMOTE-ENN, has significantly enhanced the gain and lift, hence ranking positive instances higher in order to provide actionable insights.

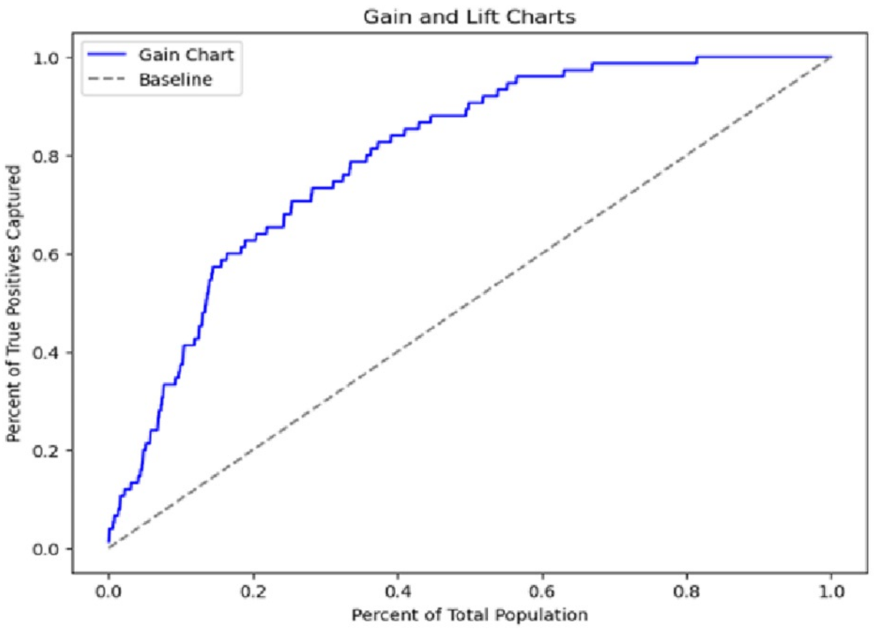

FIGURE 14: Gain and lift chart

These together provide a full evaluation of the performance of the models involved and, more importantly, the efficiency of SMOTE-ENN in bolstering the performance of LightGBM on imbalanced data. Excellent results depicted in confusion matrix, F1 score vs. threshold, calibration curve, precision-recall curve, CAP curve, learning curve, and gain and lift charts consolidate the idea of advanced resampling techniques being employed toward the construction of strong predictive models in difficult data conditions.

Results

Table *4* provides a metrics comparison (baseline vs. SMOTE-ENN with LightGBM), and Figure*15* presents an ROC curves comparison to enhance comprehension of the entire article. Table *5* presents the comparison among some models performing with SMOTE to deal with an imbalance for which accuracy is measured. Liu et al. [7] presented a model, in 2019, with an accuracy of 71.6%. Wu and Fang[8], in 2020, outperformed the previous entry with an accuracy of 78%. Butt et al. [10], in 2022, achieved 84.11% accuracy, outweighing previous tasks. A model is proposed here that, in combination with 2024 SMOTE-EN with LightGBM, greatly outperforms previous works, reaching an accuracy of 95.8%. This means that great improvements can come forth when advanced techniques are incorporated for handling imbalanced datasets. The data presented in this study are available at the following link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

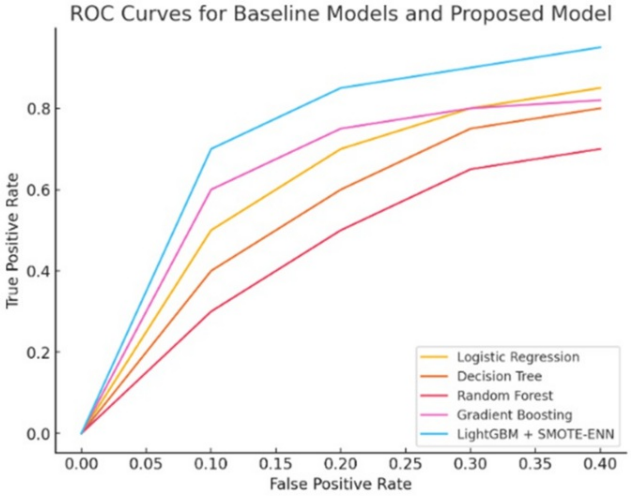

FIGURE 15: Metrics comparison using ROC curves (baseline vs. SMOTE-ENN with LightGBM)

ROC: Receiver Operating Characteristic; SMOTE-ENN: Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors; LightGBM: Light Gradient-Boosting Machine

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.746	0.161	0.758	0.266	0.854
Decision Tree	0.925	0.317	0.21	0.252	0.59
Random Forest	0.939	0	0	0	0.804
Gradient Boosting	0.94	1	0.016	0.032	0.828
LightGBM + SMOTE-ENN	0.958	0.853	0.917	0.884	0.946

TABLE 4: Metrics comparison (baseline vs. SMOTE-ENN with LightGBM)

AUC: Area Under the Curve: SMOTE-ENN: Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors; LightGBM: Light Gradient-Boosting Machine

Ref. No.	Authors	Year	Accuracy
[7] (SMOTE)	Liu et al.	2019	71.6
[8] (SMOTE)	Wu and Fang	2020	78
[10] (SMOTE)	Butt et al.	2022	84.11
[16] EMS (Elastic Net–MLP–SMOTE)	Merdas	2024	95
Our Model (SMOTE-ENN- LightGBM)		2024	95.8

TABLE 5: Performance comparison

SMOTE-ENN: Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors; LightGBM: Light Gradient-Boosting Machine

Discussion

These results and visualizations collectively highlight the challenges and solutions associated with imbalanced datasets. Initial baseline models, while efficient in their foundation in balanced scenarios, could not suffice in this context given the severe imbalance between the positive and negative classes. This is clearly reflected by the low precision, recall, and F1 scores of these models; the random forest model failed miserably in correctly classifying instances of the minority class. SMOTE application did alleviate some of these issues by increasing the recall of the models but at the cost of precision. The justification behind this was that the models became more sensitive in terms of detecting the minority class but created more false positives in the process. This trade-off demonstrated the need for a more refined approach to balance these critical metrics.

This was achieved particularly well by the SMOTE-ENN technique. By their oversampling of the minority class and removal of noisy data points through ENN, the SMOTE-ENN cleaned up and balanced the dataset. The results showed that in performance metrics and visualizations, the LightGBM model coupled with SMOTE-ENN yielded a better balance between precision and recall. The increased model AUC, combined with its performance on the precision-recall and CAP curves, underlines its efficiency in handling the imbalanced dataset. The importance of SMOTE-ENN now is that, not only was it capable of cleaning noise, but it also balanced the classes. This allowed generalization for the LightGBM model and reduced chances of overfitting to the minority class instances, hence increasing the overall robustness of the model. This enhancement is particularly pertinent in health care, where false positives and false negatives could be associated with significant consequences. Compared to the other models in the experiment, the tuned LightGBM model after SMOTE-ENN reaches an overall better balance between sensitivity and specificity, hence a reliable and efficient system that could be of great value in the field for making predictions of rare events like strokes.

Conclusions

This work demonstrated the effectiveness of the LightGBM algorithm coupled with the SMOTE-ENN technique to handle class imbalance problems in stroke prediction datasets. Our proposed approach not only boosted sensitivity and specificity in predictive models but also outperformed traditional methods, as reflected by improved AUC, precision, recall, and F1 scores. Furthermore, with RandomizedSearchCV, the model was further tuned for better hyperparameter optimization that resulted in robust and reliable forecasts. This is novel, hence setting a new standard for the models developed for stroke prediction, especially in handling imbalanced datasets as are commonly found within medical diagnosis contexts.

In the future, several ways of research are open. Testing the scalability and adaptability with bigger and more diverse datasets would consolidate the model even more across various demographic and regional boundaries. Further, other advanced machine learning algorithms and hybrid models could also be explored in the search for ever-better methods to handle class imbalances. This may extend our approaches to other medical conditions that also have challenges of data imbalance, hence extending the wide impact on healthcare. Finally, integrating real-time processing of data and embedding these models into clinical decision support systems could enable actionable insights at the point of care.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Kaliprasanna Swain

Drafting of the manuscript: Kaliprasanna Swain, Tan Kuan Tak, Kamal Upreti, Sivaneasan Bala Krishnan, Ramesh Chandra Poonia, Sumya Ranjan Nayak, Mihir Narayan Mohanty

Supervision: Kaliprasanna Swain, Tan Kuan Tak, Kamal Upreti, Pravin R. Kshirsagar, Sivaneasan Bala Krishnan, Ramesh Chandra Poonia, Sumya Ranjan Nayak, Mihir Narayan Mohanty

Acquisition, analysis, or interpretation of data: Tan Kuan Tak, Kamal Upreti, Pravin R. Kshirsagar, Sivaneasan Bala Krishnan, Ramesh Chandra Poonia, Sumya Ranjan Nayak, Mihir Narayan Mohanty

Critical review of the manuscript for important intellectual content: Pravin R. Kshirsagar

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Alanazi EM, Abdou A, Luo J: Predicting risk of stroke from lab tests using machine learning algorithms: development and evaluation of prediction models. JMIR Formative Research. 2021, 5:e23440. 10.2196/23440
2. Amann J: Machine learning in stroke medicine: opportunities and challenges for risk prediction and prevention. Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues. Advances in Neuroethics. Jotterand F, Ienca M (ed): Springer, Cham, Switzerland; 2021. 57-71. 10.1007/978-3-030-74188-4_5
3. Saceleanu VM, Toader C, Ples H, et al.: Integrative approaches in acute ischemic stroke: from symptom recognition to future innovations. Biomedicines. 2023, 11:2617. 10.3390/biomedicines11102617
4. Johnson JM, Khoshgoftaar TM: Survey on deep learning with class imbalance. Journal of Big Data. 2019, 6:27. 10.1186/s40537-019-0192-5
5. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002, 16:321-357. 10.1613/jair.953
6. Alkhawaldeh IM, Albalkhi I, Naswhan AJ: Challenges and limitations of synthetic minority oversampling techniques in machine learning. World Journal of Methodology. 2023, 13:373-378. 10.5662/wjm.v13.i5.373
7. Liu T, Fan W, Wu C: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial Intelligence in Medicine. 2019, 101:101723. 10.1016/j.artmed.2019.101723
8. Wu Y, Fang Y: Stroke prediction with machine learning methods among older Chinese. International Journal of Environmental Research and Public Health. 2020, 17:1828. 10.3390/ijerph17061828
9. Tazin T, Alam MN, Dola NN, Bari MS, Bourouis S, Khan MM: Stroke disease detection and prediction using robust learning approaches. Journal of Healthcare Engineering. 2021, 2021:7633381. 10.1155/2021/7633381
10. Butt MO, Rehman U, Javaid S, Ali TM, Nawaz A: An application of artificial intelligence for an early and effective prediction of heart failure. 2022 Third International Conference on Latest Trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, Pakistan. 2022. 1-6. 10.1109/INTELLECT55495.2022.9969182
11. Santos LI, Camargos MO, D'Angelo MFSV, Mendes JB, de Medeiros EEC, Guimarães ALS, Palhares RM: Decision tree and artificial immune systems for stroke prediction in imbalanced data. Expert Systems with Applications. 2022, 191:116221. 10.1016/j.eswa.2021.116221
12. Biswas N, Mohi Uddin KM, Rikta ST, Dey SK: A comparative analysis of machine learning classifiers for stroke prediction: a predictive analytics approach. Healthcare Analytics. 2022, 2:100116. 10.1016/j.health.2022.100116
13. Wang J, Gu H, Gu H: Optimizing stroke prediction in machine learning by addressing data imbalance. 2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM). 2023, 665-669. 10.1109/ISPCEM60569.2023.00125
14. Dahiya M, Mishra N, Agarwal S, Parveen Z: Predicting the occurrence of ischemic stroke by gradient boost approaches. 2023 4th International Conference on Intelligent Engineering and Management (ICIEM). 2023, 1-4. 10.1109/ICIEM59379.2023.10166287
15. Ushasree D, Praveen Krishna AV, Rao Ch. M: Enhanced stroke prediction using stacking methodology (ESPESM) in intelligent sensors for aiding preemptive clinical diagnosis of brain stroke. Measurement: Sensors. 2024, 33:101108. 10.1016/j.measen.2024.101108
16. Merdas HM: Elastic Net - MLP - SMOTE (EMS)-based model for enhancing stroke prediction. Medinformatics. 2024, 1:73-78. 10.47852/bonviewmedin42022470