Hybrid and Ensemble Learning Framework for Accurate Classification of Imbalanced Data

P. Sirish Kumar[1], Sai Kiran Oruganti[2]
[1,2]Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia
Email id: pdf.sirish@lincoln.edu.my

Abstract: This study evaluates advanced machine learning models for classification in imbalanced datasets. Two hybrid models, SVM–MLP and Autoencoder–XGBoost, are proposed and compared with LightGBM and XGBoost using MCC, Balanced Accuracy, and ROC–AUC. The hybrid models achieve superior performance, with SVM–MLP providing the most balanced and reliable predictions, while Autoencoder–XGBoost effectively handles class imbalance. The results demonstrate that hybrid learning frameworks combining feature extraction and powerful classifiers significantly improve classification performance in imbalanced datasets.

Keywords: Imbalanced classification, Hybrid machine learning, Support Vector Machine (SVM), Autoencoder, XGBoost.

Introduction

Classification tasks are a Classification is an essential task in machine learning and is used in applications such as healthcare, fraud detection, and predictive maintenance. However, many real-world datasets are imbalanced, where one class has much more data than the other. This causes traditional models to focus more on the majority class and often miss the minority class, which is usually more important. As a result, the predictions can become biased and less reliable [1]. To address this problem, advanced methods that combine effective feature extraction and reliable classification techniques are needed. These approaches help the model learn patterns from both classes and improve overall prediction performance. This study explores such methods to improve classification accuracy on imbalanced datasets and ensure more reliable and balanced results in real-world applications [2].

Literature Study

Imbalanced datasets continue to be a major challenge in classification, as traditional machine learning models often favor the majority class and perform poorly on the minority class [3]. To reduce this problem, data balancing methods such as SMOTE and Tomek Links are commonly used to improve minority class representation [4]. However, these methods may introduce noise or require additional preprocessing. Recently, hybrid models have shown better performance on imbalanced datasets. Autoencoders help extract important features from complex data, which improves classification accuracy [5]. Similarly, combining neural networks with methods like Support Vector Machines helps capture complex patterns and improves model generalization [6]. Ensemble methods such as Random Forest and LightGBM also improve performance by combining multiple models [7]. In addition, using proper evaluation metrics such as Matthews Correlation Coefficient and Balanced Accuracy provides a more accurate assessment of model performance on imbalanced data [8], [9]. These metrics help ensure that both classes are treated

fairly. Overall, hybrid and ensemble methods offer a reliable solution for improving classification performance in imbalanced datasets [10].

Methodology

The dataset contains 60,000 samples with 14 attributes describing an electric grid stability system. It includes time constants (tau1–tau4), generator active power values (p1–p4), and reactance values (g1–g4). A stability index (stab) and a categorical label (stabf) indicate stable or unstable grid conditions. The dataset is suitable for developing machine learning models for grid stability prediction..

Let the dataset be $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$represents the feature vector and $y_i \in \{0,1\}$the class label. In the first hybrid approach, a linear Support Vector Machine (SVM) is used for feature selection by solving $\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i$subject to $y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$. Feature importance is determined from the weight magnitudes $| w_j |$, and the top $k$ranked features form the reduced vector x_i'. These features are then used to train a Multi-Layer Perceptron (MLP), where the hidden layer representation is computed as $h^{(l)} = f(W^{(l)} h^{(l-1)} + b^{(l)})$with $h^{(0)} = x_i'$. The output layer applies softmax $P(y_i = c \mid x_i') = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}$, and the predicted class is obtained as $\hat{y}_i = \arg\max_c P(y_i = c \mid x_i')$. In the second hybrid model, an autoencoder learns compact feature representations by minimizing the reconstruction loss $L_{recon} = \frac{1}{N} \sum_{i=1}^{N} \| x_i - \hat{x}_i \|^2$, producing latent features $z_i = f_{enc}(x_i)$. These encoded features are then used to train an XGBoost classifier with objective $L_{XGB} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$, where predictions are updated iteratively as $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(z_i)$. For a new sample x_{new}, the encoded representation $z_{new} = f_{enc}(x_{new})$is obtained and the final class prediction is produced using the trained XGBoost model.

Results & Discussion

Table I shows that the hybrid models achieve the best overall performance. SVM with MLP provides the highest test accuracy (97.76%) and maintains strong precision and recall for both classes, indicating reliable and balanced predictions. Autoencoder + XGBoost also performs very well, achieving 97.73% accuracy with consistent precision and recall values, confirming the effectiveness of autoencoder-based feature extraction.

Table 1. Sample columns and signal parameters included in the GNSS dataset used in this study

Model	Validation Accuracy	Test Accuracy	Precision		Recall		F1-Score	
			Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Random Forest	0.9446	0.9386	0.94	0.93	0.96	0.9	0.95	0.91
LightGBM	0.9574	0.9584	0.96	0.95	0.97	0.93	0.97	0.94
Autoencoder + XGBoost	0.9773	0.9773	0.98	0.97	0.98	0.97	0.98	0.97
SVM with MLP	0.9784	0.9776	0.97	0.99	0.99	0.95	0.98	0.97

LightGBM achieves good performance with 95.84% accuracy and balanced results, though its recall is slightly lower for the minority class. Random Forest delivers stable performance (93.86% accuracy) but shows relatively lower recall for unstable grid conditions. Overall, hybrid models demonstrate clear advantages in handling imbalanced data.

Evaluation Metrics

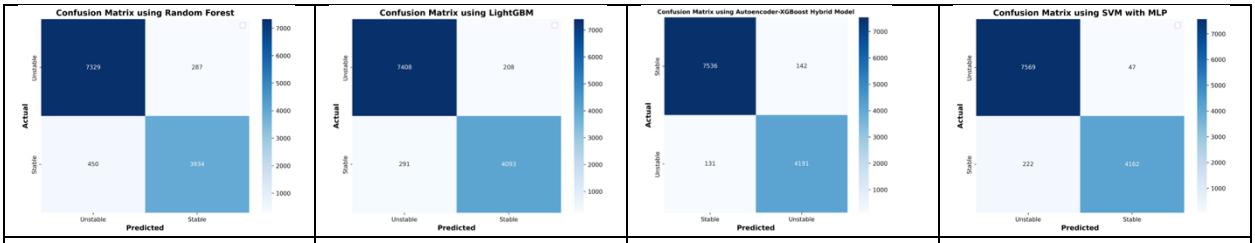
Table II presents MCC, Balanced Accuracy, and ROC AUC Score for all models. SVM with MLP achieves the highest MCC (0.9517), indicating strong agreement between predictions and actual labels. Autoencoder + XGBoost achieves the highest Balanced Accuracy (0.9756), showing effective performance across both classes. SVM with MLP also records the highest ROC AUC score (0.9985), demonstrating excellent class separation. These results confirm the effectiveness of hybrid models in imbalanced classification tasks.

Table 2. Evaluation metrics for machine learning models in grid stability Prediction

Metric	RF	LightGBM	Autoencoder + XGBoost	SVM with MLP
MCC	0.8669	0.9101	0.9507	0.9517
Balanced Accuracy	0.9298	0.9532	0.9756	0.9716
ROC AUC Score	0.9885	0.9939	0.9981	0.9985

Confusion Matrix Analysis

Figure 1 shows the confusion matrices for all models. Random Forest produces more classification errors, especially for the minority class. LightGBM improves prediction accuracy and reduces misclassification. Autoencoder + XGBoost further reduces errors and provides more balanced predictions. SVM with MLP achieves the lowest number of misclassifications, demonstrating its ability to capture complex patterns and improve prediction reliability.

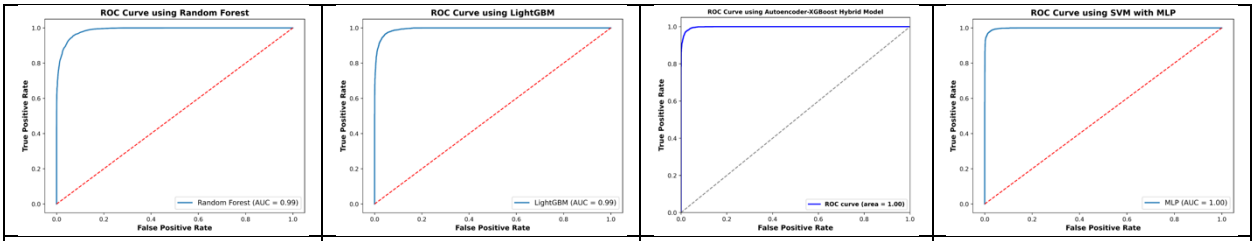


(a). RF | *(b). LightGBM* | *(c). Autoencoder + XGBoost* | *(d). SVM with MLP*

Figure 1. Confusion Matrices for Grid Stability Prediction Models

ROC Curve Analysis

Figure 2 presents the ROC curves. The hybrid models, SVM with MLP and Autoencoder + XGBoost, achieve near-perfect ROC AUC values, indicating excellent class discrimination. LightGBM and Random Forest also perform well but are slightly less effective compared to hybrid models.



(a). RF	*(b). LightGBM*	*(c). Autoencoder + XGBoost*	*(d). SVM with MLP*

Figure 2. ROC Curves for Grid Stability Prediction Models

Hybrid models outperform conventional approaches for grid stability prediction, with SVM–MLP achieving the best overall performance and Autoencoder–XGBoost also delivering strong results. These findings highlight the effectiveness of hybrid techniques for handling imbalanced classification problems..

Conclusion

Hybrid models SVM–MLP and Autoencoder–XGBoost achieved the best performance for the imbalanced dataset, delivering high accuracy and reliable predictions for both majority and minority classes. Their effectiveness highlights the advantage of combining feature selection or extraction with powerful classifiers. Overall, the results demonstrate that hybrid and ensemble approaches are highly effective for imbalanced classification, offering strong potential for reliable real-world applications and future model development.

References

1. Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y., "XGBoost: Extreme gradient boosting," R package version 1.7.6, 2023.
2. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y., "LightGBM: A highly efficient gradient boosting decision tree," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, pp. 4769–4783, 2022.
3. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. I., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
4. Buda, M., Maki, A., and Mazurowski, M. A., "A systematic study of the class imbalance problem in convolutional neural networks," Neural Networks, vol. 106, pp. 249–259, 2019.
5. Japkowicz, N., and Stephen, S., "The class imbalance problem: A systematic study," Intelligent Data Analysis, vol. 23, no. 2, pp. 429–449, 2019.
6. Goodfellow, I., Bengio, Y., and Courville, A., "Deep learning," MIT Press, Cambridge, MA, 2021.
7. Aggarwal, C. C., "Neural networks and deep learning," Springer, Cham, Switzerland, 2023.
8. Raschka, S., Patterson, J., and Nolet, C., "Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," Information, vol. 11, no. 4, pp. 193–210, 2020.
9. Chicco, D., Tötsch, N., and Jurman, G., "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy," BioData Mining, vol. 14, no. 13, pp. 1–22, 2021.
10. Shwartz-Ziv, R., and Armon, A., "Tabular data: Deep learning is not all you need," Information Fusion, vol. 81, pp. 84–90, 2022.