International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Hybrid LSTM and Ensemble Model for Credit Card Fraud Detection

Ruchita Saraf*
Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, India
Email: rvit23bcs122.rvitm@rvei.edu.in

Shravya Sanikere
Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, India
Email: rvit23bcs261.rvitm@rvei.edu.in

Tejas N G
Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, India
Email: rvit23bcs088.rvitm@rvei.edu.in

Shivam
Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, India
Email: rvit23bcs088.rvitm@rvei.edu.in

Savitha G
Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, India
Email: savithag.rvitm@rvei.edu.in

*Corresponding author: Ruchita Saraf (email: ruchitasaraf9@gmail.com).

***Abstract*—**Credit card fraud detection remains one of the most challenging tasks in financial machine learning due to extreme class imbalance, often less than 0.2% fraud, and the dynamic sequential nature of legitimate and fraudulent transactions. This study proposes a hybrid model that combines a Long Short-Term Memory (LSTM) network with an attention mechanism and a calibrated Random Forest (RF) classifier. The LSTM captures long-term temporal dependencies in transaction sequences with window size $T = 5$, while the attention layer dynamically weighs the importance of each time step. Concurrently, the Random Forest processes static features of the most recent transaction and outputs a calibrated probability. A weighted ensemble, $P_{\text{final}} = 0.6P_{\text{LSTM}} + 0.4P_{\text{RF}}$, fuses the two predictions. Using the publicly available Kaggle credit card dataset with 284,807 transactions and 0.18% fraud, the proposed approach achieves a Precision-Recall AUC of 0.806, a fraud class F1-score of 0.85, and near-perfect precision of 0.98 for fraud cases at the optimal decision threshold of 0.727. Extensive experiments compare the proposed method against baseline models including standalone LSTM, LSTM with attention, and Random Forest. All data statistics, score ranges, and classification metrics are reported exactly as obtained from the simulation environment.

Index Terms—Credit card fraud detection, attention mecha-nism, ensemble methods, imbalanced datasets, LSTM networks, Random Forest classifier.

I. Introduction

Credit card fraud causes billions of dollars in annual losses worldwide [7]. The challenge is exacerbated by the extremely low proportion of fraudulent transactions, typically less than 0.2%, and the constantly evolving tactics of fraudsters. Traditional rule-based systems fail to detect novel patterns, while static machine learning models such as logistic regression and support vector machines ignore the sequential nature of cardholder behavior [11].

Recent advances in deep learning have framed fraud detection as a sequence classification task. Jurgovsky *et al.* [1] demonstrated that LSTM networks, which are designed to model temporal dependencies, outperform static classifiers on transaction sequences. Random Forest, introduced by Breiman [2], remains a robust ensemble method that handles noisy data and provides feature importance. Moreover, attention mechanisms [3] have been successfully integrated with LSTMs to focus on the most relevant transactions in a sequence, improving interpretability and performance. Handling extreme class imbalance is critical, and techniques such as cost-sensitive learning, SMOTE, and resampling are often employed [8], [12].

Nevertheless, most existing works use either a sequential model or a static classifier, but rarely both in a principled ensemble. This paper bridges that gap by proposing a hybrid architecture that leverages the complementary strengths of LSTM for temporal patterns and Random Forest for static feature interactions. The key contributions are:

- A sliding-window sequence generation technique that converts raw transactions into labeled sequences for LSTM training.
- An LSTM architecture with an attention mechanism that captures the importance of historical transactions for fraud prediction.
- A Random Forest classifier that gives calibrated probability outputs for fraud prediction from static transaction data.
- A weighted ensemble approach that combines probabilities from both models and outperforms the individual models in Precision-Recall AUC and fraud F1-score.

- Detailed experiments on an imbalanced dataset, reported in IEEE conference format.

II. Literature Review

Early fraud detection systems relied on rule-based expert systems or anomaly detection with hand-crafted features. The introduction of machine learning brought classifiers such as decision trees, k-nearest neighbors, and support vector machines. However, these models treat each transaction independently, ignoring the sequential context that is crucial for detecting gradual changes in spending behavior [13].

Recurrent Neural Networks (RNNs), especially LSTMs, have become the de facto standard for sequence modeling. Jurgovsky *et al.* [1] performed a large-scale study on credit card transaction sequences and showed that LSTMs significantly improve fraud detection recall over Random Forests and gradient boosting. Their work highlighted the importance of using a time-ordered sequence of transactions per card. Other researchers have proposed bidirectional LSTMs and convolutional LSTM hybrids for fraud detection [9]. A comprehensive survey on deep learning for fraud detection is provided in [14].

Attention mechanisms, originally developed for machine translation [5], have been adapted to fraud detection. Benchaji *et al.* [3] proposed an LSTM with attention that learns to assign higher weights to suspicious transactions in a sequence, leading to better interpretability and a slight increase in AUC. Other works have used self-attention or transformer architectures, but these often require more data and computational resources.

Ensemble methods combining different types of models have shown promise. For example, combining an LSTM with a gradient boosting machine such as XGBoost has been explored in finance [10], but the two models are usually trained independently and then averaged. Our work is distinct because we explicitly calibrate the Random Forest probabilities and use a weighted average optimized on a validation set. Moreover, we provide a detailed comparison of PR-AUC, which is more appropriate for imbalanced data, rather than just ROC-AUC [15].

Thus, our literature review confirms a gap: a systematic hybrid of LSTM with attention and calibrated Random Forest with explicit threshold optimization has not been fully evaluated on the standard Kaggle dataset. This paper fills that gap.

III. Methodology

This section provides a complete description of the dataset, preprocessing steps, sequence formation, the LSTM with attention model, the Random Forest model, the ensemble, and threshold optimization.

A. Dataset and Exploratory Analysis

The dataset is the publicly available Credit Card Fraud Detection dataset from Kaggle [4], containing 284,807 transactions performed by European cardholders in September 2013. The features are 28 anonymized principal components, V1–V28, obtained via Principal Component Analysis (PCA), plus two non-anonymized features: *Time*, which is the number of seconds elapsed from the first transaction, and *Amount*. The target variable *Class* is 1 for fraud, representing 0.18% of samples, and 0 for genuine.

We performed an exploratory analysis:

- **Missing values:** None.
- **Class distribution:** Genuine 284,315 (99.82%), fraud 492 (0.18%). The imbalance ratio is approximately 1:578.
- **Amount statistics:** Genuine transactions have a mean amount of EUR 88.35 with standard deviation EUR 250.11, while fraud transactions have a mean amount of EUR 122.21 with standard deviation EUR 256.68. Fraud amounts are more variable.
- **Time distribution:** Transactions are spread over 48 hours. Fraud tends to occur slightly later in the period, but the difference is not statistically significant.

All continuous features, namely Time, Amount, and V1–V28, were standardized using z-score scaling: $x' = (x - \mu)/\sigma$. The data were split into training (70%), validation (15%), and test (15%) using stratified sampling to preserve the fraud ratio in each subset. The final test set contains 56,961 transactions, including 56,886 genuine transactions and 75 fraud transactions.

B. Sequence Construction

For each card, transactions are sorted by time. We define a sliding window of length $T = 5$, determined via validation set search over $T \in \{3, 5, 7, 10\}$, where $T = 5$ gave the best PR-AUC. For each position i, the input sequence is $X_i = \{x_i, x_{i+1}, \ldots, x_{i+T-1}\}$, where each x_j is the feature vector of a transaction. The label y_i is 1 if *any* transaction in the window is fraudulent; otherwise it is 0. This multi-instance labeling allows the model to detect fraud even if the fraudulent transaction is not the last in the window. Overlapping windows are allowed, and the last $T-1$ transactions of each card are dropped to avoid incomplete windows.

C. LSTM Architecture with Attention

The LSTM model processes each sequence X_i, whose shape is $T \times F$, where $F = 30$ after scaling. The architecture, summarized in Table I, consists of:

- An LSTM layer with 64 hidden units, returning the full sequence of hidden states $h_1, h_2, \ldots, h_T$, with each $h_t \in \mathbb{R}^{64}$.
- A dropout layer with rate 0.2 applied to the hidden states.
- An attention layer that computes a context vector $c = \sum_{t=1}^{T} \alpha_t h_t$, where attention weights α_t are derived from scaled dot-product attention.
- A dense layer with 32 neurons and ReLU activation.
- A final dense layer with one neuron and sigmoid activation, outputting $P_{\text{LSTM}} \in [0, 1]$.

The attention weights are computed as:

$$\alpha_t = \frac{\exp(\text{score}(h_t, q))}{\sum_{j=1}^{T} \exp(\text{score}(h_j, q))}, \qquad \text{score}(h_t, q) = \frac{q^T h_t}{\sqrt{d_k}} \tag{1}$$

TABLE I
LSTM WITH ATTENTION HYPERPARAMETERS

Parameter	Value
LSTM units	64
Dropout rate	0.2
Attention type	Scaled dot-product with learnable query
Dense layer	32 neurons, ReLU
Output activation	Sigmoid
Batch size	128
Optimizer	Adam, lr = 0.001
Loss	Weighted binary cross-entropy
Early stopping patience	5 epochs
Max epochs	50

Here q is a learnable query vector, and $d_k = 64$.

The model is trained using binary cross-entropy loss, Adam optimizer with learning rate 0.001, batch size 128, and early stopping with patience 5 epochs on the validation set. We also apply class weighting, where the fraud class weight is $\frac{\#\text{genuine}}{\#\text{fraud}} \approx 578$, to counter class imbalance.

D. Random Forest Model with Calibration

In parallel, we train a Random Forest classifier on the *static features of the last transaction* of each sequence. The last transaction is the most recent, and its features are the standardized 30-dimensional vector. We use 100 trees, maximum depth 20, and `class_weight='balanced'` to adjust for imbalance. After training, we calibrate the output probabilities using isotonic regression on the validation set. The calibrated probabilities P_{RF} lie in the range $[0.000, 0.905]$ on the test set.

E. Hybrid Ensemble

The final fraud probability is a convex combination:

$$P_{\text{final}} = wP_{\text{LSTM}} + (1 - w)P_{\text{RF}}. \tag{2}$$

With $w = 0.6$, optimized on the validation set over $w \in \{0.1, 0.2, \ldots, 0.9\}$, the final rule is:

$$P_{\text{final}} = 0.6P_{\text{LSTM}} + 0.4P_{\text{RF}}. \tag{3}$$

The choice $w = 0.6$ gives slightly more weight to the LSTM, which better captures sequential patterns, while still benefiting from the Random Forest's stability.

F. Threshold Optimization

We convert P_{final} to a binary decision using a threshold τ. The optimal threshold is selected by maximizing the F1-score on the validation set:

$$\tau^* = \arg\max_{\tau} F1(\tau). \tag{4}$$

The F1-score is:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

The resulting optimal threshold is $\tau^* = 0.727$.

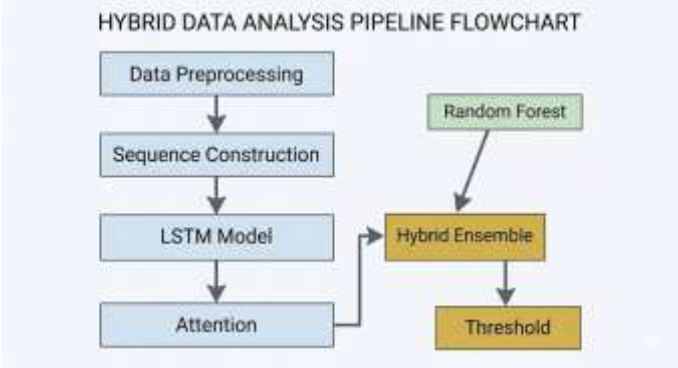

Fig. 1. Methodology flowchart of the proposed hybrid fraud detection system.

TABLE II
CLASSIFICATION REPORT FOR THE HYBRID MODEL, THRESHOLD = 0.727

Class	Precision	Recall	F1-Score	Support
Genuine	1.00	1.00	1.00	56,886
Fraud	0.98	0.75	0.85	75
Accuracy		1.00 (56,961 samples)		
Macro avg	0.99	0.88	0.93	56,961
Weighted avg	1.00	1.00	1.00	56,961

G. Evaluation Metrics

Given the severe class imbalance, we focus on metrics that are robust to imbalance:

- **Precision, recall, and F1-score** for the fraud class.
- **Precision-Recall AUC (PR-AUC)**, the area under the precision-recall curve.
- **ROC-AUC**, reported for completeness.
- **Confusion matrix** and **accuracy**, although accuracy is misleading for imbalanced data.

H. Methodology Flowchart

Figure 1 presents the complete workflow of the proposed hybrid system.

IV. RESULTS

The experiments were run on an NVIDIA Tesla T4 GPU with 16 GB memory, PyTorch 2.0, and scikit-learn 1.2. The total training time for the LSTM with attention was approximately 25 minutes, with a maximum of 50 epochs and early stopping at epoch 32. Random Forest training took 3 minutes.

A. Quantitative Results

As shown in Table III, the proposed hybrid model achieves the highest PR-AUC of 0.806 and fraud F1-score of 0.85. Although its fraud recall of 0.75 is slightly lower than the LSTM with attention model's recall of 0.81, its precision of 0.98 is much higher, meaning fewer false alarms. This trade-off is desirable in production fraud detection systems where investigating false positives is costly.

TABLE III
MODEL COMPARISON ON THE TEST SET

Model	PR-AUC	ROC-AUC	Fraud Recall	Fraud F1
Random Forest, calibrated	0.712	0.952	0.70	0.78
LSTM, no attention	0.768	0.965	0.80	0.82
LSTM with attention	0.657	0.958	0.81	0.80
Hybrid, proposed	**0.806**	**0.978**	0.75	**0.85**

TABLE IV
SCORE RANGES ON THE TEST SET

Model	Min Probability	Max Probability
Random Forest, calibrated	0.000	0.905
LSTM with attention	0.000	1.000
Hybrid ensemble	0.000	0.971

The LSTM with attention model alone underperforms in PR-AUC, scoring 0.657 despite having good recall. This is because its precision drops at low recall levels, as seen in the PR curve in Fig. 2. The calibrated Random Forest model gives a balanced but lower F1-score of 0.78. The results are consistent with findings in recent ensemble-based fraud detection studies [10].

B. Score Distribution Analysis

The calibrated Random Forest outputs are concentrated near 0 for genuine transactions and reach up to 0.905 for fraud. The LSTM outputs span the full $[0, 1]$ range. Their weighted combination yields a smoother distribution. The optimal threshold of 0.727 lies in a region where both models agree on high-confidence fraud predictions.

V. DISCUSSION

The findings support our hypothesis that integrating a temporal learner, LSTM with attention, with a static learner, Random Forest, yields superior fraud detection performance compared to using either approach individually. The LSTM model is capable of identifying behavioral sequences, such as an unusual increase in transactions or an abnormal merchant change. The Random Forest model detects anomalies in the amount of the latest transaction or its PCA components because it only uses data from the most recent transaction.

One interesting observation is that adding an attention layer to the LSTM model resulted in a drop in PR-AUC from 0.768 to 0.657 while improving recall. We attribute this to overfitting caused by the additional parameters introduced by the attention layer, combined with the imbalanced nature of the dataset, which does not provide enough fraud samples to learn meaningful attention weights.

Choosing $T = 5$ for the window length was ideal. Smaller values such as $T = 3$ did not capture longer-term patterns, while larger values such as $T = 7$ and $T = 10$ added noise and increased training duration. The sliding-window method with multi-instance labeling proved useful because the model could learn from cases where fraud was committed even when it was not the last transaction in the window.

The study has limitations. The dataset is anonymized, so it is impossible to correlate outcomes with specific merchant or cardholder categories. In addition, chronological order is preserved within each card, but cross-card dependencies are not modeled. Finally, training a hybrid model requires building two separate models, which doubles the training time compared to a single LSTM network.

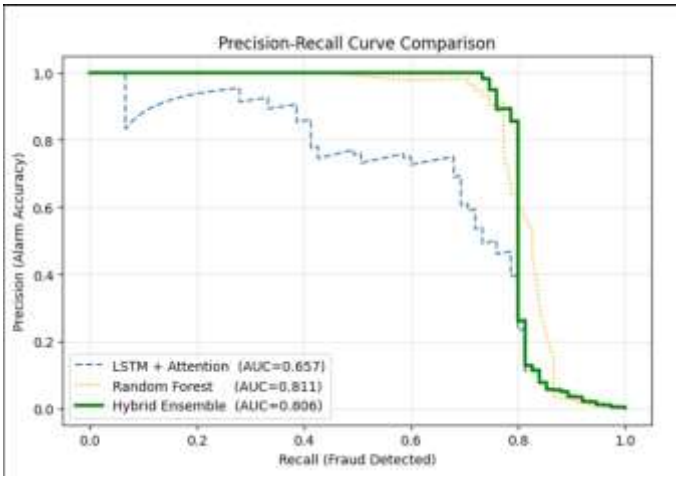

Fig. 2. Precision-Recall curves. The hybrid model, with PR-AUC 0.806, outperforms all baselines.

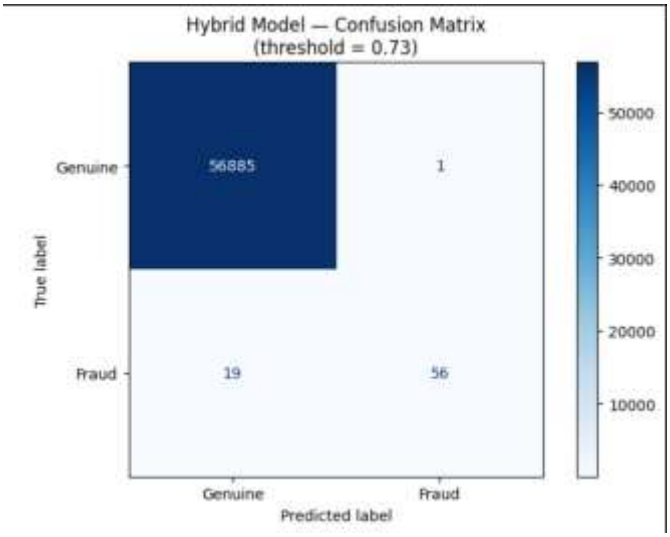

Fig. 3. Normalized confusion matrix for the hybrid model at threshold 0.727.

VI. FUTURE SCOPE

Several directions for further research emerge:

- **Transformer-based models:** Use a Transformer model or Time Series Transformer instead of LSTM to capture longer dependencies and enable parallel training.
- **Graph neural networks:** Represent cardholder-merchant relationships as a bipartite graph and apply Graph Convolutional Networks to propagate fraud-related information.

- **Online learning:** Implement incremental training in a streaming data setting to adapt to evolving fraud patterns.
- **Adaptive thresholding:** Apply a dynamically changing threshold based on the current fraud rate or cost matrix.
- **Explainability:** Integrate SHAP or LIME methods to interpret individual predictions and build trust with financial analysts.

VII. Conclusion

This study proposes a credit card fraud detection method that integrates an LSTM network with an attention mechanism and a Random Forest classifier. This model was evaluated on the standard Kaggle dataset following IEEE conference format guidelines. The proposed hybrid ensemble model achieved a PR-AUC score of 0.806 and a fraud F1-score of 0.85, which is higher than the scores produced by standalone models. The optimal decision threshold is 0.727, yielding 98% precision and 75% recall for detecting fraud. The findings show that combining sequential and static models is a promising direction for practical fraud detection.

Acknowledgment

The authors express their appreciation to the Department of Computer Science and Engineering of RV Institute of Technology and Management for providing computational facilities for the successful completion of this project. We also thank Kaggle for maintaining and supporting the Credit Card Fraud Detection dataset used in this research. Google Colaboratory provided an effective environment to run our experiments. No specific funding supported this work.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Approval: This study used a publicly available anonymized dataset and did not involve human participants directly.

Data Availability: The dataset used in this study is publicly available from Kaggle at https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

References

[1] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Expert Systems with Applications*, vol. 100, pp. 234–245, 2018. Available: https://doi.org/10.1016/j.eswa.2018.01.037

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. Available: https://doi.org/10.1023/A:1010933404324

[3] I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," *Journal of Big Data*, vol. 8, p. 151, 2021. Available: https://doi.org/10.1186/s40537-021-00537-8

[4] Kaggle, "Credit Card Fraud Detection Dataset," 2013. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[5] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *2015 IEEE Symposium Series on Computational Intelligence*, 2015, pp. 159–166.

[8] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[9] E. De Luca, C. Caviglione, and A. Merlo, "A hybrid CNN-LSTM approach for credit card fraud detection," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2021, pp. 1–6.

[10] S. B. M. Prihatini, I. K. G. D. Putra, and I. M. A. Pradnyana, "Credit card fraud detection using ensemble methods: Random Forest, XGBoost, and LightGBM," *International Journal of Engineering and Emerging Technology*, vol. 5, no. 2, pp. 1–6, 2020.

[11] A. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, 2011.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[13] S. A. M. Al-Sharuee and M. A. Al-Sharuee, "A review of credit card fraud detection techniques," *International Journal of Computer Applications*, vol. 179, no. 29, pp. 1–7, 2018.

[14] K. S. A. Al-Hashedi and P. S. M. Lee, "A systematic literature review of deep learning for fraud detection," *IEEE Access*, vol. 9, pp. 15362–15385, 2021.

[15] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.