

Evaluation of Deep Learning Models for Tabular Data Classification of Deposit Customers

Gilang Miftakhul Fahmi^{1,*}, Fajar Rohmattulloh², Gupita Nurmalita Sari³, Dewi Kartika Putri⁴, Iqbaluddin Syam Had⁵

^{1,2,3,4,5}, Fakultas Ilmu Komputer, Program Studi Magister Ilmu Komputer, Universitas Amikom Purwokerto, Indonesia

e-mail: ^{1,*}gilangfahmi20@gmail.com, ²fajarrohmat1@gmail.com, ³gupitasari96@gmail.com,

⁴dewip9419@gmail.com, ⁵iqbaluddinsh@gmail.com

Article History	Abstract
Received : 06 January 2026 Revised : 22 February 2026 Accepted : 02 March 2026 Published : 28 March 2026 Keywords: <i>deep learning, tabular data, deposit classification, convolutional neural network, banking.</i>	The application of deep learning to tabular data faces challenges in the banking domain, which is dominated by numerical and categorical features without a temporal structure. This study aims to evaluate and compare the performance of several deep learning architectures for classifying time deposit customers. The models used include Multilayer Perceptron, Convolutional Neural Network, and Long Short-Term Memory-based Recurrent Neural Network. The dataset consists of banking customer data with a binary classification target: subscribing or not subscribing to deposits. The research stages included data preprocessing, model development, training, and performance evaluation using accuracy, precision, recall, and area under the ROC curve metrics. Analysis performed on the training curve to assess optimization stability and generalization ability. The experimental results showed that all models achieved an accuracy above 89 percent. The Convolutional Neural Network model provides most balanced performance with the highest accuracy and precision as well as stable training dynamics. Multilayer Perceptron shows good class separation capabilities but tends to overfit, while Recurrent Neural Network does not provide significant performance improvements and shows instability during training. These findings confirm the suitability of deep learning architecture to the characteristics of tabular data is an important factor in model selection for banking applications.

*)Corresponding author : dewip9419@gmail.com

1. INTRODUCTION

The rapid development of deep learning has brought significant changes to various machine learning tasks, particularly in processing unstructured data such as images and text. The primary advantage of this approach lies in its ability to automatically learn complex feature representations without heavy reliance on manual feature engineering (Goodfellow et al., 2016). This success has encouraged the adoption of deep learning in other domains, including business and finance, which are predominantly characterized by structured tabular data.

Tabular data remains the most commonly used data format in real-world applications, especially in the banking, insurance, and data-driven marketing sectors. Such data typically consists of a mixture of numerical and categorical features, imbalanced class distributions, and feature relationships that are not necessarily linear. Recent studies have shown that these characteristics often prevent conventional deep learning models from consistently outperforming classical machine learning methods, such as tree-based models [1].

Several comprehensive studies report that tree-based ensemble algorithms, including Random Forest and XGBoost, frequently outperform neural networks on tabular data, particularly when the dataset size is limited and the feature structure is heterogeneous [2]. These findings challenge the

assumption that deep learning is a universal solution for all types of data and emphasize that model selection must carefully consider the intrinsic characteristics of the dataset.

The limitations of deep learning on tabular data are also associated with the absence of explicit spatial or sequential structure among features, making it difficult for neural networks to effectively capture local interactions between attributes. Other common issues include a tendency toward overfitting, complex hyperparameter tuning requirements, and sensitivity to data preprocessing techniques [3]. These conditions indicate that the success of deep learning on tabular data is highly dependent on architectural design choices and training strategies.

Several approaches have been proposed to address these limitations, including the use of Convolutional Neural Networks by transforming tabular data into specific representations that allow the exploitation of local feature relationships. Other approaches apply Recurrent Neural Network architectures by treating features as pseudo-sequences; however, the effectiveness of this strategy remains debatable, as tabular data inherently lacks a natural temporal structure [4].

This study focuses on comparing the performance of several commonly used deep learning models namely Multilayer Perceptron, Convolutional Neural Network, and Long Short-Term Memory on the task of bank deposit customer classification. The study not only evaluates model performance using quantitative metrics but also analyzes training stability and generalization capability across architectures. This approach is expected to provide deeper insights into the suitability of specific deep learning models for tabular data in the banking domain. Based on a review of prior research, a research gap remains in the empirical comparative analysis of deep learning architectures applied to banking tabular data, particularly with respect to training dynamics and indications of overfitting. Therefore, the objective of this study is to evaluate and compare the performance of MLP, CNN, and LSTM models in deposit classification tasks and to identify the architectural characteristics most appropriate for tabular data without explicit temporal structure.

2. METHOD

2.1 Research Stages

This study adopts a quantitative experimental approach aimed at evaluating the performance of several deep learning architectures for tabular data classification tasks. The research stages are systematically designed to ensure that each model is tested under equivalent data conditions and evaluation scenarios, thereby enabling fair and objective performance comparisons. The research begins with data collection and an initial understanding of the banking customer dataset used as the object of study. The dataset then undergoes a preprocessing stage to ensure data

quality and input format compatibility with deep learning models. The subsequent stage involves model architecture design, training procedures, and performance evaluation using relevant metrics. The final stage focuses on the comparative analysis of experimental results and the interpretation of model behavior based on the characteristics of tabular data. The design of these research stages follows common practices in applied machine learning model evaluation, where experimental consistency is a critical factor in maintaining the validity and reliability of research findings [5].

2.2 Dataset and Data Characteristics

The dataset used in this study consists of banking customer data intended to predict customers' propensity to subscribe to term deposit products. The target variable is binary, representing customers who subscribe to deposits and those who do not. The available attributes include demographic information, occupational details, and financial characteristics represented through a combination of numerical and categorical features.

The characteristics of this dataset reflect common properties of tabular data in the banking domain, namely heterogeneous feature types and an imbalanced class distribution. These conditions make the dataset a relevant case study for evaluating the effectiveness of various deep learning architectures on structured, non-temporal data [1]. The dataset is divided into training and testing subsets

using an 80:20 ratio. This split is applied to assess the model's generalization capability on unseen data that is not involved during the training process.

2.3 Data Preprocessing

The data preprocessing stage is conducted to ensure that all features can be effectively processed by deep learning models. Categorical features are transformed into numerical representations using encoding techniques to ensure compatibility with mathematical operations within neural networks. One-hot encoding is employed to avoid imposing artificial ordinal assumptions on categorical variables, while label encoding is selectively applied to features with a large number of categories.

All numerical features are subsequently normalized using standardization techniques to align feature scales across attributes. Standardization aims to accelerate training convergence and reduce the dominance of features with large value ranges on the loss function [6],[7],[8]. The preprocessing procedures are consistently applied to both training and testing data to prevent information leakage.

2.4 Deep Learning Model Architectures

This study evaluates three deep learning architectures with distinct characteristics in learning data representations: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network based on Long Short-Term Memory (LSTM). The Multilayer Perceptron model is employed as a baseline approach, as it represents a standard feedforward neural network commonly applied to tabular data. MLP consists of multiple hidden layers with nonlinear activation functions designed to capture complex relationships among features. This architecture is frequently used as a baseline in evaluating deep learning models on structured data [2].

The Convolutional Neural Network model is implemented using one-dimensional convolution operations to capture local patterns among features. This approach is based on the assumption that local interactions between subsets of features may provide additional discriminative information for classification. Several studies have shown that CNNs can improve performance on certain tabular datasets when input representations are appropriately designed [9]. The Recurrent Neural Network model based on LSTM is used to evaluate the capability of sequential networks in modeling dependencies among features treated as pseudo-sequences. Although tabular data lacks inherent temporal properties, this approach is examined to analyze the impact of sequential assumptions on training stability and classification performance [4].

2.5 Training and Evaluation Process

All models are trained using the binary cross-entropy loss function and the Adam optimization algorithm due to its stability across various neural network training scenarios. The training process is conducted using the same number of epochs for each model to ensure direct comparability of experimental results. Model performance is evaluated using multiple metrics, including accuracy, precision, recall, and the area under the Receiver Operating Characteristic (ROC) curve. The use of multiple evaluation metrics aims to provide a more comprehensive assessment of model performance, particularly under imbalanced class distribution conditions. This multi-metric evaluation approach aligns with current recommendations in applied classification system evaluation research [8].

3. RESULTS AND DISCUSSION

3.1 Model Evaluation Results

Model performance was evaluated using accuracy, precision, recall, and the area under the ROC curve to provide a comprehensive assessment of each model's classification capability. The use of multiple evaluation metrics is necessary because accuracy alone is not sufficiently representative in binary classification tasks with potential class imbalance [13],[14],[15].

Table 1. Evaluation Results

Model	Accuracy	Precision	Recall	ROC-AUC
MLP	89.09%	0.54	0.29	0.63

CNN	89.86%	0.64	0.24	0.61
RNN	89.11%	0.55	0.25	0.61

The experimental results indicate that all three deep learning models achieved relatively high accuracy levels, exceeding 89%. The Convolutional Neural Network (CNN) obtained the highest accuracy at 89.86%, followed by the Recurrent Neural Network (RNN) at 89.11% and the Multilayer Perceptron (MLP) at 89.09%. The relatively small differences in accuracy among the models suggest that result interpretation should consider additional metrics that are more sensitive to prediction distribution. The highest precision value was achieved by the CNN model at 0.64, while the MLP and RNN models reached precision values of 0.54 and 0.55, respectively. Higher precision indicates that CNN is more consistent in correctly identifying customers who truly subscribe to deposit products compared to the other models. This finding is particularly relevant in banking applications, as false positive predictions may directly affect the efficiency and cost-effectiveness of marketing strategies.

Recall values across all models were in the low to moderate range, with MLP achieving 0.29, CNN 0.24, and RNN 0.25. The relatively low recall values indicate that the models still struggle to capture all customers who genuinely have the potential to subscribe to deposit products. This phenomenon commonly occurs in datasets with imbalanced class distributions, where models tend to prioritize the majority class. The highest ROC AUC value was obtained by the MLP model at 0.63, while the CNN and RNN models both achieved a value of 0.61. The higher ROC AUC score of MLP suggests superior global class separation capability, even though it does not consistently produce highly precise positive predictions. These results emphasize that model performance cannot be adequately evaluated using a single metric alone.

3.2 Comparative Analysis of Model Architectures

The superior performance of CNN in terms of accuracy and precision suggests that one-dimensional convolution operations are more effective in capturing local interactions among tabular features compared to purely fully connected architectures. This finding aligns with previous studies reporting that CNNs can model correlations between neighboring features when input representations are consistently designed [10],[11],[12].

The Multilayer Perceptron model demonstrates competitive performance, particularly with respect to the ROC AUC metric. This indicates that MLP is capable of learning an overall decision boundary reasonably well but is less effective in producing stable positive predictions. This limitation may be attributed to the tendency of MLP to overfit the training data, especially when the number of model parameters is relatively large compared to the dataset size [2].

The Recurrent Neural Network based on LSTM does not exhibit significant performance improvements over MLP and CNN. This result supports the view that applying RNN architectures to tabular data without inherent temporal structure offers limited benefits. Treating features as pseudo-sequences may introduce inappropriate sequential assumptions that increase optimization complexity without contributing meaningful information.

3.3 Training Dynamics and Generalization Analysis

Analysis of the training curves reveals clear indications of overfitting in the MLP model. Training accuracy increases consistently, while validation accuracy tends to stagnate and decline during later epochs. This pattern is further supported by increasing validation loss despite a continuous decrease in training loss, indicating that the model excessively adapts to the training data at the expense of generalization capability.

The CNN model exhibits the most stable training dynamics among the three architectures. The accuracy and loss curves for training and validation data follow similar trends with relatively small gaps. This stability suggests that CNN is able to extract relevant patterns without excessively memorizing the training data. Such characteristics strengthen the argument that CNN is more suitable for certain tabular datasets compared to the other architectures evaluated [1].

The RNN model demonstrates less stable training behavior, characterized by significant fluctuations in training loss across several epochs. These loss spikes indicate optimization instability, which is often associated with issues such as exploding gradients or suboptimal learning rate

configurations in sequential networks. This condition further reinforces the finding that RNN architectures require additional handling to be effectively applied to non-sequential data.



Figure 1. Visualization of Training Results

3.4 Implications of the Research Findings

The results of this study indicate that the selection of deep learning architectures for tabular data must consider the alignment between model structure and data characteristics. The CNN demonstrates the best balance between predictive performance and training stability in the bank deposit customer classification task. The MLP remains relevant as a baseline model with strong class separation capability, while the application of RNNs to tabular data without an explicit temporal dimension should be critically evaluated. These findings provide practical implications for the development of decision support systems in the banking sector, particularly in selecting deep learning-based predictive models that are both efficient and stable. The comparative approach employed in this study can also serve as a foundation for future research exploring more specialized deep learning architectures tailored for tabular data.

4. CONCLUSION

This study evaluated the performance of three deep learning architectures Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network based on Long Short-Term Memory (LSTM) for bank deposit customer classification using tabular banking data. The experimental results indicate that all models achieved relatively high accuracy levels; however, notable differences were observed in performance characteristics and training stability across the three architectures. The Convolutional Neural Network demonstrated the most balanced performance, achieving the highest accuracy and precision while exhibiting stable training dynamics. These findings suggest that one-dimensional convolution operations are effective in capturing local interactions among

tabular features without inducing dominant overfitting. The Multilayer Perceptron showed reasonably strong class separation capability based on ROC AUC values but exhibited a more pronounced tendency toward overfitting, leading to reduced generalization performance in the later stages of training. The Recurrent Neural Network did not yield meaningful performance improvements and showed training instability, reinforcing the argument that sequential architectures are less suitable for tabular data without inherent temporal structure. The limitations of this study include the use of a single type of tabular dataset and the absence of direct comparisons with classical machine learning methods. These constraints limit the generalizability of the findings to other tabular domains with different characteristics. Future research is encouraged to incorporate datasets with more diverse scales and class distributions, as well as to compare deep learning approaches with tree-based methods or architectures specifically designed for tabular data. Further exploration of regularization techniques and architectural adaptations also holds potential for improving model stability and performance in structured, data-driven banking applications.

BIBLIOGRAPHY

- [1] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data," in *Advances in Neural Information Processing Systems*, 2022.
- [2] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, 2022.
- [3] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Networks Learn. Syst.*, 2021.
- [4] A. Kadra, M. Lindauer, F. Hutter, and J. Grabocka, "Tabular data modeling with attention and deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
- [5] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2021.
- [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [7] T. Salimans and D. P. Kingma, "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [8] J. Kohler, H. Daneshmand, A. Lucchi, T. Hofmann, M. Zhou, and K. Neymeyr, "Exponential Convergence Rates for Batch Normalization: The Power of Length-Direction Decoupling in Non-convex Optimization," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 89, pp. 806–815, 2019.
- [9] X. Huang, J. Zhu, X. Li, and L. Wang, "Convolutional neural networks for tabular data classification," *Neurocomputing*, 2020.
- [10] H. A. Alenizy and J. Berri, "Transforming Tabular Data into Images via Enhanced Spatial Relationships for CNN Processing," *Scientific Reports*, vol. 15, Art. no. 17004, 2025, doi: 10.1038/s41598-025-01568-0.
- [11] Z. Wang, M. Zheng, and P. X. Liu, "Robust and Interpretable Tabular Data Classification via Multi-Channel Image Conversion and Channel-Wise Gating," *Neural Networks*, vol. 205, Art. no. 109316, 2026, doi: 10.1016/j.neunet.2026.109316.
- [12] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot," *PLoS One*, 2015.
- [13] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 820–836, 2024, doi: 10.1162/TACL_A_00675.
- [14] H. Ali, M. S. Shamsuddin, and A. L. Ralescu, "Imbalanced Classification: A Paradigm-Based Review," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 361–386, 2022, doi: 10.1002/sam.11538.

-
- [15] A. Chiu et al., "Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review," *Brain Sciences*, vol. 13, no. 12, Art. no. 1723, 2023.