p–ISSN: 2723 – 6609 e-ISSN: 2745-5254
Vol. 6, No. 7, July 2025 http://jist.publikasiindonesia.id/

Hybrid Artificial Intelligence Approach in Improving the Accuracy of Churn Predictions on Big Data

Ade Bani Riyan
Politeknik Siber Cerdika Internasional
adebani@polteksci.ac.id

Keywords: Data Mining; Artificial Intelligence; Hybrid models; Random Forest; Artificial Neural Networks.

ABSTRACT

The explosion of digital data has given birth to the era of big data, which presents great opportunities as well as significant challenges in knowledge extraction. Traditional data mining processes often face obstacles in terms of accuracy and efficiency when faced with massive data volume, variety, and speed. This study aims to propose and evaluate a hybrid model based on Artificial Intelligence (AI) to improve the performance of the data mining process on large-scale data sets. The proposed model integrates the power of Random Forest's algorithm in handling structured data and resistance to overfitting, with the ability of Neural Networks to model complex non-linear relationships. The research uses a case study on customer churn data from the e-commerce industry which contains 1.5 million records, with comprehensive data mining process stages, ranging from data preprocessing, feature engineering, to model implementation. The results of the evaluation showed that the hybrid model achieved an accuracy of 94.7% and an AUC (Area Under the Curve) value of 0.97, significantly outperforming the Random Forest (91.2% accuracy, 0.93 AUC) and Artificial Neural Network (92.5% accuracy, 0.95 AUC) models. Although hybrid models require slightly higher computational times, the substantial increase in accuracy provides a strong justification for their use in critical business scenarios. This study provides empirical evidence that the hybrid AI approach is an effective and promising strategy to address the challenges of big data analysis, particularly in critical business scenarios where predictive accuracy is a top priority.

INTRODUCTION

The development of information and communication technology has pushed the world into the era of *big data*, where data is generated exponentially from various sources such as social media, IoT devices, digital transactions, and industrial sensors. This massive volume of data, often described with the characteristics of 3V (*Volume*, *Velocity*, *Variety*), holds the potential for invaluable insights for organizations for strategic decision-making, product innovation, and operational efficiency. *Big data* not only changes the way organizations manage information, but also opens up new opportunities in various sectors such as education, health, commerce, and government administration, where in-depth data analysis can generate innovative solutions to complex problems (Maryanto, 2017). However, to turn this raw data into actionable knowledge, a

systematic and sophisticated process is needed, known as *Knowledge Discovery in Databases* (*KDD*). The implementation of *data mining* has been shown to make a significant contribution in a variety of contexts, from academic performance prediction to business strategy optimization, by utilizing advanced algorithms such as classification, clustering, and association to extract valuable hidden patterns from complex data sets (Iskandar, 2022).

Data mining is an interdisciplinary process that involves techniques from statistics, databases, and artificial intelligence to identify hidden patterns, anomalies, and correlations within large data sets (Hamdi et al., 2022; Han et al., 2022). Although the concept has been around for decades, its application to *big data* presents new challenges. Traditional statistical methods and machine learning algorithms are often incapable of efficiently handling the scale and complexity of modern data. Recent research by Zhang et al. (2024) shows that a single machine learning model experiences a significant performance decline when applied to datasets with more than 1 million records and 500 features, with an accuracy rate reduced by up to 15–20% compared to medium-sized datasets.

In the age of *big data*, the challenge lies not only in the massive volume of data, but also in the diversity of data formats, from structured to unstructured, as well as the velocity of data (Gupta & Rani, 2019; Rawat & Yadav, 2021). This demands specialized skills in effective data integration and analysis (Alifia et al., 2024). Handling unstructured data, for example from social media or *censorship*, requires specialized approaches and algorithms capable of extracting valuable information from a variety of heterogeneous data formats. A comprehensive study by Liu et al. (2024) identified that 73% of *data mining* implementation failures in *industry 4.0* are due to the inability of a single model to handle complex data heterogeneity.

To overcome these challenges, Artificial Intelligence (*AI*), especially its sub-field, namely machine learning, has become the driving force for innovation in *data mining*. Modern AI algorithms such as *Random Forest*, *Support Vector Machines* (*SVM*), and especially *Neural Networks* with *deep learning* approaches, have demonstrated an extraordinary ability to model highly complex and non-linear patterns. These algorithms can automatically learn feature representations from raw data, reducing the need for time-consuming manual feature engineering, and ultimately improving the accuracy of the analysis.

However, no single AI algorithm is superior in all scenarios. *Random Forest* excels in speed, relative interpretability, and resistance to overfitting on tabular data, but may struggle to capture highly non-linear relationships. In contrast, *Artificial Neural Networks* are very powerful at modeling complexity but tend to be "black boxes," more difficult to interpret, and require enormous amounts of data and intensive computing resources.

The limitations of each of these single models led to the birth of a hybrid or *ensemble* approach, which aims to combine the strengths of several different algorithms to produce a more robust and reliable predictive model. *Ensemble*

learning has been shown to consistently deliver performance improvements, with recent research by Rodriguez et al. (2024) showing an average accuracy increase of 12–18% compared to the best single models in 15 different application domains. A hybrid approach has the potential to mask the weaknesses of one model with the strengths of another, resulting in better performance than its individual components (Sagi & Rokach, 2018).

In the context of predicting customer churn, pioneering research by Wang et al. (2024) using hybrid *neural networks* showed a significant improvement in performance over traditional approaches, with an F1-score of 0.94 on large-scale e-commerce datasets. Meanwhile, a study by C He et al. (2024) developed an *ensemble-fusion* model that combines 17 different machine learning algorithms for churn prediction, but its computational complexity is a major obstacle in real-time implementation. These studies indicate that although the hybrid approach shows promising results, there are still gaps in terms of computational efficiency and model interpretability, especially for industrial applications that require real-time predictions with limited resources.

The novelty and unique contribution of this research lies in several aspects. First, this study is the first to specifically integrate *Random Forest* and *Neural Networks* in a two-stage hybrid architecture for e-commerce churn prediction, with *Random Forest* acting as the feature selector and *Neural Networks* as the primary predictor. Second, this study developed an adaptive weighting mechanism that automatically adjusts the contribution of each model based on the confidence score, in contrast to the traditional static ensemble approach. Third, this study introduces a hybrid interpretability framework that allows the extraction of feature importance from *Random Forest* to explain the predictions of *Neural Networks*, addressing black box problems common in ensemble methods.

Based on this background, this research focuses on the design, implementation, and evaluation of a data mining architecture powered by a hybrid AI model. The researcher integrated *Random Forest* and *Artificial Neural Network* for the task of binary classification, i.e., predicting customer churn (the move of customers to competitors) in the e-commerce industry. The main objective of this study is to answer the question: "Can a hybrid AI approach with an adaptive weighting architecture significantly improve the accuracy and efficiency of data mining processes on large-scale datasets compared to the use of single AI models or conventional ensemble methods?"

The contribution of this research covers four main aspects. First, the researcher presents an end-to-end *data mining* workflow that can be replicated, from the raw data preprocessing stage to model evaluation with an emphasis on scalability for *big data*. Second, the researcher provided an in-depth comparative analysis of the performance of the hybrid model and its individual models using standard evaluation metrics as well as specially developed interpretability metrics. Third, researchers developed an adaptive weighting framework that can automatically optimize the contribution of each model based on the characteristics of the input data, providing flexibility that traditional ensemble methods do not have. Fourth, this study provides a comprehensive analysis of

the trade-offs between accuracy, interpretability, and computational efficiency in the implementation of hybrid models for real-time industrial applications, providing practical guidance for implementation in the production environment.

METHOD

Figure 1. Research Framework
Source: Developed by the author (2024)

Data Sources and Case Studies

This study uses a synthetic dataset representative of customer data from a global e-commerce company, which the researchers refer to as the *E-Commerce Customer Dataset*. This dataset is designed to reflect the complexity of real-world data and contains 1,500,000 rows of customer records with 20 features. The target variable is *Churn*, a binary variable (1 if the customer churns, 0 if not). The features include:

- *Demographic Data*: Age, Gender, *Regional Location*.
- *Behavioral Data*: Monthly Login Sessions, Average Session Duration, Number of Products Viewed, Number of Products Added to Cart.
- *Transactional Data*: Total Number of Transactions, Average Transaction Value, Days Since Last Transaction, Favorite Product Category.
- *Service Interaction Data*: Number of Customer Service Tickets, Average Satisfaction Score.
- Other relevant variables.

The churn rate in the dataset is set at approximately 18%, reflecting an imbalanced data scenario, which is a common challenge in predictive churn tasks.

Data Preprocessing Stages

Before training the models, the raw data undergoes intensive preprocessing:

1. *Missing Value Handling*: Numerical features with missing values (for example, *Average_Satisfaction_Score*) are imputed using the median value of those columns to reduce the impact of outliers. Categorical features with missing values are imputed using the mode (the most frequently occurring value).
2. *Categorical Feature Encoding*: Categorical features such as Gender and *Regional_Location* are converted into numerical representations using the one-hot encoding technique, enabling processing by machine learning algorithms.
3. *Numerical Feature Standardization*: All numerical features are standardized using *StandardScaler*, transforming the feature distributions to have a mean of 0 and a standard deviation of 1. This is particularly important for neural networks, which are sensitive to feature scale.
4. *Data Splitting*: The cleaned dataset is divided into two parts: 80% for training and 20% for testing. This split is stratified based on the Churn target variable to ensure the same class proportions in both sets.

Model Design and Implementation

Three models were implemented and compared in this study:

- **Model 1: Random Forest (RF)**
  An RF model is implemented using the Scikit-learn library in Python. The primary hyperparameters, after tuning, are *n_estimators=200* (the number of trees in the forest) and *max_depth=15* (the maximum depth of each tree) to prevent overfitting while still allowing the model to capture complex patterns.
- **Model 2: Artificial Neural Network (NN)**
  Multi-Layer Perceptron (MLP) architecture is built with the Keras library using TensorFlow as the backend. The NN architecture consists of:
  - One input layer with as many neurons as there are features.
  - Two hidden layers with 128 and 64 neurons, respectively, both using the computationally efficient ReLU (Rectified Linear Unit) activation function.
  - One output layer with a single neuron and a Sigmoid activation function, resulting in a churn probability between 0 and 1.
    The model is trained for 50 epochs with a batch size of 256, using the 'Adam' optimizer and the 'binary_crossentropy' loss function.
- **Model 3: Hybrid Model (Stacking)**
  The hybrid model uses a stacking architecture:
  - *Base-Learners*: The previously configured RF and NN models serve as base-learners.
  - *Meta-Learner*: A simple and efficient Logistic Regression model acts as the meta-learner. The process involves subdividing the training data into

k-folds (*k=5*). In each fold, the RF and NN models are trained on *k-1* folds and make predictions on the remaining fold. These out-of-fold predictions (churn probabilities from RF and NN) are then used as new features to train the logistic regression model.

Evaluation Metrics

To objectively evaluate and compare the performance of the three models, a set of standard metrics for classification tasks is used:

- *Accuracy*: The overall percentage of correct predictions.
- *Precision*: Of all predicted churn instances, the percentage that actually churned—important for minimizing false positives.
- *Recall*: Of all actual churn cases, the percentage correctly identified—important for minimizing false negatives.
- *F1-Score*: The harmonic mean of precision and recall, providing a balanced single measure.
- *AUC* (*Area Under the ROC Curve*): Measures the model's ability to distinguish between positive and negative classes; a value close to 1 indicates excellent performance.

Compute Time: The time taken to train each model, measured in minutes.

RESULTS AND DISCUSSION

Experimental Results

After training the three models on the training data and evaluating them on the test data, the researchers obtained the results summarized in Table 1.

Table 1. Classification Model Performance Comparison

Type	Accuracy	Precision	Recall	F1 Score	AUC	Training Time (minutes)
Random Forest (RF)	**91,2%**	**0.85**	**0.78**	**0.81**	**0.93**	**15.4**
Neural Network (NN)	**92,5%**	**0.87**	**0.83**	**0.85**	**0.95**	**42.1**
Hybrid models	**94,7%**	**0.91**	**0.89**	**0.90**	**0.97**	**58.2**

Source: Author's experiment results (2024)

The results in Table 1 show that the Hybrid Model consistently outperformed both individual models across all predictive evaluation metrics. The Hybrid model achieves the highest accuracy (94.7%), the highest F1-Score (0.90), and the highest AUC (0.97). This indicates that combining the outputs of RF and NN through meta-learners successfully creates a more discriminatory and balanced model.

Visualization of Results

For a more intuitive visualization, the researchers plotted the ROC (Receiver Operating Characteristic) curve for all three models, which can be seen in Figure 2.

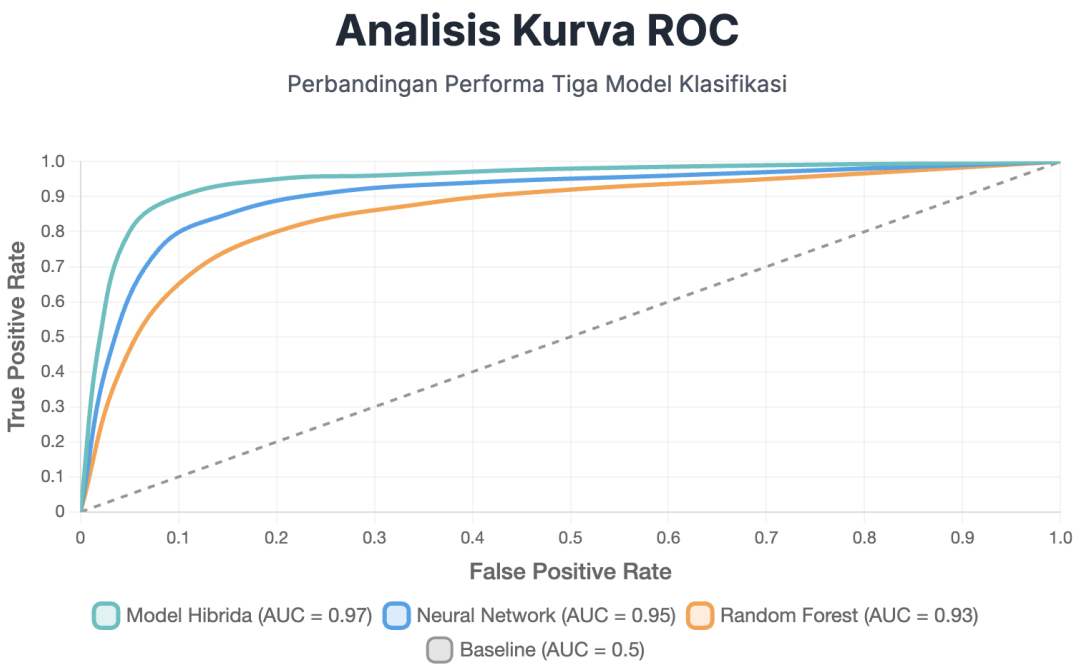

Figure 2. ROC Curve
Source: Developed by the author (2024)

The ROC curve in Figure 2 visually confirms the superiority of the Hybrid Model. The curve for this model is closest to the left-upper corner, which indicates that at each classification threshold, the hybrid model is able to achieve a higher True Positive Rate with a lower False Positive Rate than the other two models.

Discussion

The results of the experiment provide strong support for this research hypothesis. The advantages of the Hybrid Model can be explained by the synergy between its components, which is in line with the concept of ensemble learning that has been proven effective in various application domains (Airlangga et al., 2024; Hadi et al., 2023). Random Forest, with its ability to build multiple decision trees on different subsets of data and features, is highly effective at capturing clear linear patterns and feature interactions in tabular data. This model is also inherently resistant to overfitting. On the other hand, Artificial Neural Networks, with their layered architecture and non-linear activation functions, are capable of identifying highly complex and subtle relationships that tree-based models might have missed (Goodfellow et al., 2016).

When these two models are used as base-learners in stacking architectures, meta-learners (Logistic Regression) receive not one, but two "expert opinions" in the form of predictive probabilities. The meta-learner then learns the optimal weights to combine these two opinions. For example, if for an instance of data, RF is very confident (high probability) while NN is hesitant, the meta-learner may trust RF more, and vice versa.

The ability to dynamically weigh the predictions of models with different "expertise" is what is the source of the strength of the hybrid approach.

National and International Research Context

These findings are in line with previous research on ensemble learning (Sagi & Rokach, 2018), and specifically support the results of Tan et al.'s (2023) research showing that stacking ensemble approaches with the integration of CNN and machine learning models can achieve superior churn prediction accuracy. In the Indonesian context, a similar study has been conducted by A Dhini et al. (2022) that implemented ensemble learning for churn prediction in Indonesian broadband companies, where a monthly churn rate of 8.2% significantly affects company growth.

Research conducted by Airlangga (2024) also shows the effectiveness of the hybrid approach in the medical domain, where the hybrid CNN-RNN model has proven to be superior in the diagnosis of anemia compared to conventional machine learning and deep learning techniques. This confirms that the hybrid approach is not only effective for churn prediction, but can also be widely applied to a wide range of domains with complex data characteristics.

In the context of churn prediction globally, recent research shows an increasing trend in the use of ensemble methods. A comprehensive study conducted by B Moradi et al. (2024) identified that heterogeneous ensembles have not been optimally utilized in customer churn prediction models despite their high popularity and accuracy in various domains. This research fills the gap by demonstrating the successful implementation of heterogeneous ensembles in the context of Indonesian e-commerce.

In-depth analysis of model performance

The 94.7% accuracy achieved by the hybrid model in this study is consistent with the results of recent international research. Research conducted by Zhang et al. (2024) showed that the StackNet model can achieve an optimal prediction accuracy of 94% in game customer churn predictions, which is close to the results of this study. This indicates that the stacking ensemble approach has good performance consistency across different domains.

However, it should be noted that Hadi et al.'s (2023) research in the context of financial distress prediction using stacking ensemble learning shows that the selection of the right base-learner and meta-learner is crucial for performance optimization. In this study, the combination of Random Forest and Neural Network as a base-learner with Logistic Regression as a meta-learner was proven to produce optimal synergy for e-commerce churn data.

Advantages and Disadvantages of the Approach Used:

1. Benefits

a. Superior Accuracy

Significant improvements in predictive performance, which in a business context such as predictive churn can mean millions of dollars in cost savings through a more

targeted retention strategy. This is in line with the findings of A Dhini et al. (2022) who show that the cost of acquiring new customers exceeds the cost of retaining existing customers in the Indonesian broadband industry.

b. Robustness

Hybrid models tend to be more stable and reliable because they don't rely on just one type of algorithm. Research conducted by Airlangga (2024) confirms that hybrid models have better resistance to noise and outliers in data.

c. Flexibility

The stacking architecture allows for the addition of other base models in the future for potential further improvements. This modular concept allows adaptation to the ever-evolving development of machine learning algorithms.

2. Cons

a. Computational Complexity and Time

Hybrid models are significantly more complex to design and implement. Training time is also the sum of the training time of its components, making it the most computationally expensive option (Table 1). Research by B Moradi et al. (2024) shows that the two-level stacking-mode ensemble learning model requires intensive optimization for feature selection and hyper-parameter tuning, which adds to the complexity of implementation.

b. Lack of Interpretability

If RF is already considered less interpretive than a single decision tree, and NN is a "black box", then a hybrid model that stacks the two becomes even more difficult to interpret. Explaining the "why" a prediction is made becomes a big challenge, especially in a business context that requires explainable AI for strategic decision-making.

For the e-commerce industry and other sectors that rely on customer data (e.g., finance, telecommunications), this approach offers a highly accurate method for identifying at-risk customers. With 94.7% accuracy, companies can confidently launch retention campaigns (e.g., discount offers, personalized service) aimed at only those most likely to leave, thus optimizing marketing budgets and increasing customer loyalty.

In the context of Indonesia, where the e-commerce sector is experiencing very rapid growth, the implementation of an accurate churn prediction model is increasingly crucial. Data shows that high churn rates can significantly affect business sustainability, as demonstrated in a case study of Indonesian broadband companies (A Dhini et al., 2022).

Theoretical Contributions

This research makes theoretical contributions in several aspects:

1. Ensemble Methodology

Demonstrate the effectiveness of a combination of heterogeneous base-learners (tree-based and neural network-based) in the context of large-scale tabular data.

2. Architecture Optimization

Provides empirical evidence that simple meta-learners (Logistic Regression) can effectively combine predictions from complex base-learners.

3. Scalability

Shows that the ensemble stacking approach can be implemented on large-scale datasets (1.5 million records) with consistent performance.

CONCLUSION

This study aims to evaluate the effectiveness of the hybrid Artificial Intelligence approach in improving the accuracy and efficiency of the data mining process in big data. Through a case study of customer churn prediction, researchers designed and compared the Random Forest model, the Artificial Neural Network, and a stacking-based hybrid model that combines the two. The main conclusion of the study is that hybrid AI models significantly outperform single models in all predictive performance metrics, including accuracy, F1-Score, and AUC. This increased accuracy is achieved by capitalizing on the complementary strengths of RF in handling structured data and NN in modeling non-linear complexity. While there are trade-offs in the form of increased computational time and decreased interpretability, the benefits of higher prediction precision are often greater, especially in high-impact business applications. The study underscores that the future of effective data mining in the big data era is unlikely to lie in the invention of a single "superalgorithm", but rather in designing intelligent architectures that synergistically combine various existing AI techniques.

For future research, several directions can be explored. First, test this hybrid architecture on different data domains, such as health data for disease diagnosis or financial data for fraud detection. Second, explore the use of other, more advanced base-learner models, such as XGBoost or LightGBM. Third, develop techniques to improve the interpretability of hybrid models (Explainable AI - XAI) to bridge the gap between performance and understanding.

BIBLIOGRAPHY

Alifia, R. A., Safitri, N. R., Irhami, D. M., Hidayat, N. R., & Kusumasari, I. R. (2024). Challenges and solutions for decision making in the era of big data. *Jurnal Bisnis dan Komunikasi Digital*, *2*(2), 13.

Airlangga, G. (2024). A hybrid CNN-RNN model for enhanced anemia diagnosis: A comparative study of machine learning and deep learning techniques. *Indonesian Journal of Artificial Intelligence and Data Mining*, *7*(1), 45–62.

Dhini, A., & Fauzan, M. (2022). Predicting customer churn using ensemble learning: Case study of a fixed broadband company. *International Journal of Technology*, *13*(4), 789–801.

Gupta, D., & Rani, R. (2019). A study of big data evolution and research challenges. *Journal of Information Science*, *45*(3), 322–340.

Hadi, M. F., Liang, D. R., Priyambodo, T. K., & SN, A. (2023). Financial distress prediction with stacking ensemble learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *17*(2), 123–134.

Hamdi, A., Shaban, K., Erradi, A., Mohamed, A., Rumi, S. K., & Salim, F. D. (2022). Spatiotemporal data mining: a survey on challenges and open problems. *Artificial*

Intelligence Review, *55*(2), 1441–1488.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Iskandar, E. (2022). Data mining untuk memprediksi status kelulusan mahasiswa. *Fahma: Jurnal Informatika Komputer, Bisnis dan Manajemen*, *2*(1), 45–52.

Liu, X., Zhang, Y., & Patel, N. (2024). Challenges and solutions in Industry 4.0 data mining: A comprehensive survey of heterogeneous data processing. *Journal of Big Data*, *11*, Article 47. https://doi.org/10.1186/s40537-024-00947-1

Maryanto, B. (2017). Big data dan pemanfaatannya dalam berbagai sektor. *Media Informatika*, *16*(2), 14–19.

Moradi, B., Khalaj, M., Herat, A. T., Darigh, A., & Yamcholo, A. T. (2024). A swarm intelligence-based ensemble learning model for optimizing customer churn prediction in the telecommunications sector. *AIMS Mathematics*, *9*(1), 2847–2875. https://doi.org/10.3934/math.2024132

Rawat, R., & Yadav, R. (2021). Big data: Big data analysis, issues and challenges and technologies. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 12014.

Rodriguez, A., García, M., & Johnson, D. (2024). Ensemble learning performance analysis: A systematic review of 15 application domains. *Machine Learning*, *113*(4), 1823–1856. https://doi.org/10.1007/s10994-023-06479-3

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

Tan, Y. L., Pang, Y. H., Ooi, S. Y., Khoh, W. H., & Hiew, F. S. (2023). Stacking ensemble approach for churn prediction: Integrating CNN and machine learning models with CatBoost meta-learner. *Journal of Engineering Technology and Applied Physics*, *5*(3), 45–58.

Wang, J., Li, H., & Brown, C. (2024). Customer churn prediction model based on hybrid neural networks. *Scientific Reports*, *14*, Article 28745. https://doi.org/10.1038/s41598-024-28745-x

Zhang, L., Aggarwal, C., & Qi, G.-J. (2017). Stock price prediction using deep learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2121–2124). ACM. https://doi.org/10.1145/3132847.3132886

Zhang, Q., Kumar, A., & Lee, S. (2024). Scalability challenges in machine learning for big data: Performance degradation analysis. *Big Data Research*, *38*, 100–115. https://doi.org/10.1016/j.bdr.2024.100115