Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

A Hybrid Ensemble Approach for Enhanced Fraud Detection: Leveraging Stacking Classifiers to Improve Accuracy in Financial Transaction

Gregorius Airlangga

Engineering Faculty, Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Correspondence Author Email: gregorius.airlangga@atmajaya.ac.id
Submitted: 23/08/2024; Accepted: 31/08/2024; Published: 31/08/2024

Abstract−Fraud detection in financial transactions presents a significant challenge due to the evolving tactics of fraudsters and the inherent imbalance in datasets, where fraudulent activities are rare compared to legitimate transactions. This study proposes a Hybrid Model utilizing a stacking ensemble technique that combines multiple machines learning algorithms, including Random Forest, Gradient Boosting, SVM, LightGBM, and XGBoost, to enhance the accuracy of fraud detection systems. The Hybrid Model is evaluated against traditional machine learning models using a comprehensive cross-validation approach, with results indicating a near-perfect accuracy of 99.99%, outperforming all individual models. The study also examines the trade-offs associated with the Hybrid Model, including increased computational demands and reduced interpretability, highlighting the need for careful consideration when deploying such models in real-world scenarios. Despite these challenges, the Hybrid Model's ability to significantly reduce both false positives and false negatives makes it a powerful tool for financial institutions aiming to mitigate the risks associated with fraudulent activities. In conclusion, the findings demonstrate the effectiveness of hybrid ensemble methods in fraud detection, providing a robust solution that balances the complexities of real-world applications with the need for high accuracy. The research underscores the potential of advanced machine learning techniques in enhancing the security and reliability of financial transactions, offering valuable insights for the development of future fraud detection systems.

Keywords: Hybrid Ensemble; Fraud Detection; Stacking Classifier; Financial Transaction; Machine Learning

1. INTRODUCTION

In recent years, the proliferation of digital transactions has led to a significant increase in fraudulent activities, posing substantial risks to financial institutions and consumers alike. The need for robust, accurate, and efficient fraud detection systems has never been more urgent [1]–[3]. Traditional models, while effective to some extent, often fall short when faced with the complexity and evolving nature of fraudulent behaviors. To address these challenges, the integration of advanced machine learning techniques, particularly hybrid models, has emerged as a promising solution [4]. This research explores the application of a hybrid model, utilizing a StackingClassifier, and compares its performance with traditional models such as RandomForestClassifier, GradientBoostingClassifier, SVC, LGBMClassifier, and XGBClassifier, within the context of credit fraud detection. Fraud detection is a critical component in the financial sector, where the rapid increase in online transactions has introduced new opportunities for fraudsters to exploit system vulnerabilities [5]. Early fraud detection methods relied heavily on rule-based systems, which, despite their initial success, struggled to adapt to the dynamic nature of fraud patterns. Machine learning models have since become the cornerstone of modern fraud detection systems, offering the ability to learn from data and improve detection accuracy over time [6]–[8]. However, no single model has proven to be universally effective across all scenarios. This limitation has led to the exploration of hybrid models, which combine the strengths of multiple algorithms to achieve superior performance.

The literature on fraud detection using machine learning is extensive, reflecting the importance and complexity of the problem [9]. Traditional models like RandomForest, GradientBoosting, and Support Vector Machines have been widely studied and implemented due to their ability to handle large datasets and complex decision boundaries [10]. However, these models often require extensive feature engineering and can suffer from issues such as overfitting and high computational costs. In contrast, newer approaches, such as those involving hybrid models, offer a more flexible and scalable solution by leveraging the complementary strengths of different algorithms [11]. A hybrid model, in the context of this research, refers to the combination of multiple machine learning models in a layered structure, where the output of several base models serves as input to a meta-model. This approach, known as stacking, allows the model to capture a broader range of patterns in the data, potentially leading to improved performance [12]. Stacking has been successfully applied in various domains, including image classification, natural language processing, and financial forecasting, but its application in fraud detection, particularly in the context of credit card transactions, is still relatively underexplored.

The urgency of developing effective fraud detection models cannot be overstated. As digital payment systems become increasingly ubiquitous, the potential for financial losses due to fraud also rises [13]–[15]. According to recent studies, global losses due to credit card fraud alone are expected to exceed $30 billion annually by 2025 [16]–[18]. This staggering figure highlights the need for continued research and innovation in fraud detection techniques. Furthermore, the rapid evolution of fraudulent tactics necessitates models that can not only detect known patterns but also adapt to new, previously unseen forms of fraud [19], [20]. Current state-of-the-art methods in fraud detection primarily focus on improving the accuracy and efficiency of machine learning models.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

Techniques such as deep learning, anomaly detection, and ensemble methods have shown promise in enhancing model performance [21], [22]. However, these methods are not without their limitations. Deep learning models, for instance, require large amounts of labeled data and significant computational resources, making them impractical for some applications. Anomaly detection methods, while effective at identifying unusual patterns, often struggle with high false-positive rates. Ensemble methods, including stacking, offer a middle ground by combining the strengths of multiple models, but they require careful tuning and validation to achieve optimal results [23]

This research aims to address the gap in the literature by providing a comprehensive evaluation of a hybrid model, specifically a StackingClassifier, in the context of fraud detection. The primary goal is to determine whether the hybrid model can outperform traditional models in terms of accuracy, robustness, and scalability. The study will also explore the potential trade-offs involved in using a hybrid model, such as increased computational complexity and the need for extensive model validation. The contribution of this research lies in the development and evaluation of a novel hybrid model that combines several well-established machine learning algorithms. By comparing the performance of the hybrid model with that of traditional models, this study seeks to provide insights into the advantages and limitations of hybrid approaches in fraud detection. The results are expected to inform both researchers and practitioners about the potential benefits of adopting hybrid models in their fraud detection systems. The research methodology is structured as follows. First, the dataset used for this study is introduced, with a detailed explanation of the preprocessing steps, including imputation, scaling, and encoding of categorical variables. The choice of base models and the meta-model for the stacking classifier is then justified based on their performance in previous studies. Next, the experimental setup, including cross-validation and performance metrics, is described. The results of the experiments are presented and discussed, highlighting the performance of the hybrid model relative to traditional models. Finally, the paper concludes with a discussion of the implications of the findings, potential limitations of the study, and directions for future research.

2. RESEARCH METHODOLOGY

As presented in the figure 1, the research methodology of this study is designed to rigorously evaluate the performance of a hybrid machine learning model, specifically a StackingClassifier, in detecting fraudulent activities within credit card transactions. This section outlines the dataset, preprocessing techniques, model selection, experimental setup, and evaluation metrics used to assess the efficacy of the proposed hybrid model.

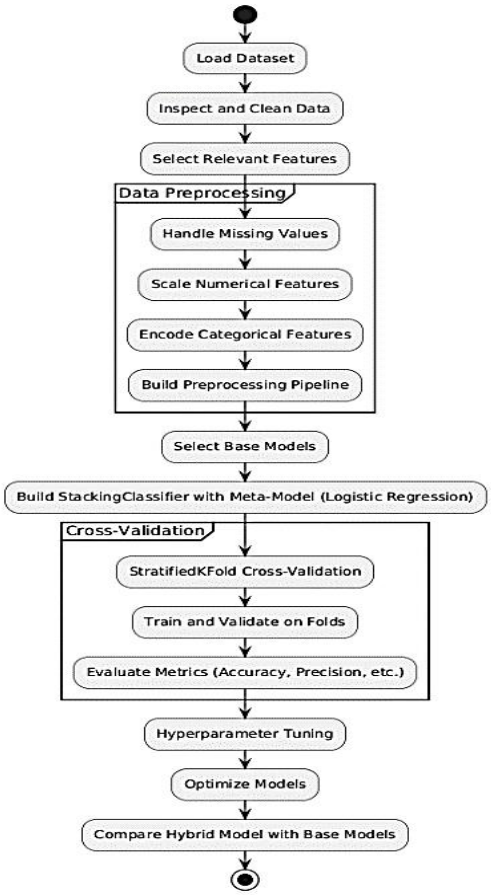

Figure 1. Research Methodology

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

The process begins by loading the dataset, which is the foundational step in any data analysis task. Once the dataset is loaded, the next step involves inspecting and cleaning the data to ensure it is free from errors, anomalies, or inconsistencies. This stage is critical as it sets the stage for accurate modeling. After cleaning the data, relevant features are selected, which involves identifying the most significant variables that will contribute to the prediction task. This step helps in reducing the complexity of the model and improving its performance. Data preprocessing follows, where several key actions are performed to prepare the data for modeling. This includes handling missing values, which might involve imputing missing data or removing incomplete records. Scaling numerical features is another essential step, ensuring that all numerical inputs are on a comparable scale, which is particularly important for algorithms sensitive to feature magnitudes. Additionally, categorical features are encoded into numerical formats that machine learning models can process effectively. These preprocessing steps are often automated through a preprocessing pipeline, ensuring consistent application across both training and testing datasets.

The next phase is the selection of base models that will form the foundation of the StackingClassifier. These base models are individually trained, and their predictions are later combined. The hybrid model is constructed by using a meta-model, in this case, Logistic Regression, which is trained on the outputs of the base models to make the final prediction. This ensemble method aims to leverage the strengths of multiple models to enhance predictive performance. To ensure the model's robustness and generalizability, cross-validation is employed. Specifically, StratifiedKFold cross-validation is used, which involves dividing the dataset into several folds while maintaining the distribution of the target variable across these folds. This method allows for thorough validation as the model is trained and validated on different subsets of the data. Throughout this process, various performance metrics such as accuracy, precision, and recall are evaluated to assess the model's effectiveness.

Following cross-validation, hyperparameter tuning is conducted to optimize the model's performance. This involves adjusting the model's hyperparameters through techniques such as GridSearchCV or RandomizedSearchCV to find the best configuration. The models are then retrained with the optimized parameters, fine-tuning them to achieve the best possible results. Finally, the performance of the hybrid model is compared with that of the individual base models. This comparison helps to determine whether the ensemble method, using a StackingClassifier with a Logistic Regression meta-model, has provided a significant improvement over the base models. The entire pipeline is designed to build a robust and accurate predictive model by combining multiple machine learning algorithms, validating their performance rigorously, and optimizing them to achieve the best outcomes.

2.1 Dataset

The dataset utilized in this study is a comprehensive collection of transaction data from a financial institution, specifically curated to detect fraudulent activities. The dataset contains transactions made by credit cards in September 2013 by European cardholders and presents transactions that occurred in two days, where it has 492 frauds out of 284,807 transactions [24]. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. The dataset's primary purpose is to serve as a foundation for training and testing various machine learning models, including the hybrid model proposed in this research. Each record within the dataset corresponds to a financial transaction and is accompanied by a variety of features that describe different aspects of the transaction. The dataset includes both numerical and categorical features that capture the multifaceted nature of financial transactions. Some of the critical features include the transaction amount, which quantifies the monetary value of the transaction; merchant details, which provide information about the entity involved in the transaction; and cardholder information, which offers insights into the profile of the individual conducting the transaction. The target variable in the dataset is binary, indicating whether a transaction is fraudulent (1) or legitimate (0). This binary classification task aligns with the study's goal of distinguishing between fraudulent and non-fraudulent transactions using machine learning models.

Before any analysis or model training could begin, a thorough examination of the dataset was conducted to ensure that only relevant features were retained. During this initial inspection, several features were identified as irrelevant or redundant for the purposes of this study. For example, the 'trans_date_trans_time' feature, which records the exact timestamp of each transaction, while valuable for time-series analysis, was deemed unnecessary for this model's design and was therefore removed. Similarly, personal details such as 'dob' (date of birth), 'first' and 'last' names, 'trans_num' (transaction number), and 'cc_num' (credit card number) were excluded. These features were removed primarily to protect privacy and because they do not contribute meaningful information to the predictive models. The dataset was then divided into two primary categories: numerical and categorical variables. Numerical variables typically include features like the transaction amount and other quantitative measures, while categorical variables consist of attributes such as the merchant's name or transaction type. This categorization is crucial for the subsequent data preprocessing steps, as it allows for the application of appropriate transformations tailored to each type of data. The decision to remove certain features and focus on the remaining ones was driven by both practical considerations and the need to streamline the dataset for efficient model training. Irrelevant features can introduce noise into the model, leading to reduced performance and increased computational complexity. By carefully selecting only the most pertinent features, the study aims to enhance the accuracy and efficiency of the models while maintaining a robust and generalizable approach to fraud detection.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

2.2 Data Preprocessing

Data preprocessing is a pivotal phase in any machine learning project, particularly in fraud detection, where the quality of the input data directly influences the performance of the models. In this study, a series of preprocessing steps were meticulously applied to prepare the dataset for model training. These steps ensured that the data was clean, well-structured, and ready for the application of various machine learning algorithms. One of the first challenges addressed during preprocessing was the handling of missing values. Missing data is a common issue in real-world datasets and can arise due to various reasons, such as incomplete data entry or system errors during data collection. If not properly addressed, missing values can lead to biased model predictions and reduce the overall reliability of the results. In this study, different strategies were employed to handle missing values based on the type of feature. For numerical features, missing values were imputed using the mean value of each respective feature. This method, known as mean imputation, is effective in preserving the central tendency of the data without introducing significant distortions. By replacing missing values with the mean, the dataset remains consistent, and the overall distribution of the numerical features is maintained. For categorical features, the most frequent value (mode) within each category was used for imputation. This approach helps preserve the categorical distributions and ensures that the imputed values are representative of the most common occurrences within the dataset.

After addressing missing values, the next step involved scaling the numerical features. Scaling is a crucial preprocessing step, particularly for models that are sensitive to the scale of the input features, such as Support Vector Machines (SVC) and Logistic Regression. Inconsistent scales can lead to models placing undue importance on features with larger ranges, skewing the predictions. To avoid this issue, numerical features were standardized using the StandardScaler. This transformation rescales the features to have a mean of zero and a standard deviation of one, ensuring that each feature contributes equally to the model's learning process. Standardization also helps in accelerating the convergence of gradient descent-based algorithms, leading to more efficient model training. Furthermore, Categorical features in the dataset were processed using an encoding technique to convert them into a numerical format suitable for machine learning models. In this study, OrdinalEncoder was employed, which transforms categorical values into integer representations. This method is particularly effective when the categorical variables have an inherent order or ranking, which is often the case in financial data. By encoding the categories into integers, the ordinal relationships are preserved, allowing the models to learn from the relative positioning of different categories.

The encoding process is crucial because machine learning models generally require numerical inputs. Without encoding, categorical features would be ignored or cause errors during model training. By converting these features into a format that the models can process, the study ensures that all relevant information is utilized, improving the overall predictive performance. All these preprocessing steps were implemented within a ColumnTransformer, a powerful tool in the scikit-learn library that allows for the systematic application of different transformations to different subsets of features. The ColumnTransformer enables the simultaneous processing of numerical and categorical data, applying the appropriate transformations to each type. This streamlined approach ensures consistency across the dataset and simplifies the preprocessing pipeline, making it easier to reproduce and validate the results.

2.3 Model Selection

The selection of models in this study is crucial to the success of the proposed hybrid approach, as it directly influences the performance and robustness of the fraud detection system. The methodology centers around the use of a StackingClassifier, a sophisticated ensemble technique that combines the strengths of multiple base models, enhancing predictive performance through diversity and complementary capabilities. StackingClassifier is a form of ensemble learning where multiple models, known as base learners, are trained independently. Their predictions are then combined and passed to a meta-learner, which makes the final prediction. This method leverages the unique strengths of each base model, allowing the ensemble to outperform any individual model by reducing biases, variances, and model-specific errors. The choice of base models and the meta-model is informed by their proven effectiveness in previous research and their complementary characteristics, ensuring that the ensemble model is both robust and versatile.

RandomForestClassifier is selected as one of the base models due to its inherent robustness and versatility. Random Forest is an ensemble of decision trees, where each tree is trained on a random subset of the data. The predictions from all trees are then averaged to produce the final output. This process reduces overfitting, a common issue in decision trees, and improves generalization to unseen data. RandomForest is particularly effective in handling large datasets with numerous features and is resilient to noise in the data, making it a strong candidate for inclusion in the stacking ensemble. GradientBoostingClassifier is another base model chosen for its ability to enhance performance through sequential learning. Unlike Random Forest, which builds trees independently, Gradient Boosting builds trees sequentially, where each new tree focuses on correcting the errors made by the previous trees. This method of boosting weak learners into strong ones makes Gradient Boosting highly effective, especially in complex datasets where capturing subtle patterns is necessary. Gradient Boosting has been widely used in fraud detection due to its high accuracy and ability to handle imbalanced datasets, which are common in this domain. Support Vector Classifier (SVC) adds a different dimension to the stacking ensemble by focusing on

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

maximizing the margin between classes in a high-dimensional space. SVC constructs hyperplanes that best separate the different classes, and its kernel trick allows it to work effectively in non-linear spaces. SVC is particularly useful when the data points are not linearly separable, as it can transform the input space into a higher-dimensional space where a clear separation is possible. Its inclusion in the ensemble introduces a model with a strong theoretical foundation and the ability to handle complex classification tasks with high precision.

LightGBM (LGBMClassifier) is included in the ensemble for its efficiency and scalability. LightGBM is a gradient boosting framework that is optimized for speed and performance. It handles large datasets with high-dimensional features more efficiently than traditional gradient boosting methods. LightGBM achieves this through a novel technique of leaf-wise tree growth, which focuses on growing the leaf with the maximum loss reduction. This approach not only speeds up the learning process but also leads to more accurate models. Given the large and complex nature of the dataset in this study, LightGBM's ability to efficiently process data while maintaining high accuracy makes it a valuable addition to the stacking ensemble. XGBoost (XGBClassifier), another variant of gradient boosting, is known for its speed and performance. XGBoost has gained popularity due to its scalability, accuracy, and flexibility. It includes a number of advanced features such as regularization, which prevents overfitting, and parallel processing, which speeds up computation. XGBoost is particularly effective in structured data tasks, making it an excellent choice for fraud detection, where the relationships between features can be complex and nonlinear. Its inclusion in the stacking ensemble brings a well-rounded model that complements the strengths of the other models.

The meta-model chosen for the stacking classifier is Logistic Regression. The role of the meta-model is to combine the predictions from the base models and produce the final output. Logistic Regression is chosen for its simplicity and effectiveness in binary classification tasks, which aligns with the objective of fraud detection. As a linear model, Logistic Regression is less prone to overfitting, especially when used in conjunction with the diverse set of base models in the ensemble. It effectively weighs the predictions of the base models, learning from their combined outputs to make accurate final predictions. Its simplicity also ensures that the ensemble remains interpretable, providing insights into how the different models contribute to the final decision. The selection of these specific models for the stacking ensemble is based on their proven track records in the domain of fraud detection, as well as their complementary strengths. By combining models that approach the classification task from different perspectives such as tree-based models, margin maximization, and gradient boosting the stacking ensemble can capture a wide range of patterns in the data, leading to superior performance in detecting fraudulent transactions.

2.4 Experimental Setup

The experimental setup is designed to rigorously evaluate the performance of the StackingClassifier in detecting fraudulent transactions. Given the high stakes associated with fraud detection, it is essential to ensure that the model not only performs well on the training data but also generalizes effectively to unseen data. To achieve this, a cross-validation approach is employed, specifically using StratifiedKFold. Furthermore, StratifiedKFold cross-validation is chosen because it preserves the class distribution in each fold, ensuring that the training and validation sets are representative of the overall dataset. This is particularly important in fraud detection, where the dataset is typically imbalanced, with a much smaller number of fraudulent transactions compared to legitimate ones. By maintaining the same proportion of fraud and non-fraud cases in each fold, StratifiedKFold reduces the risk of biased performance estimates that could arise from imbalanced class distributions.

In this setup, the dataset is split into five folds. The model is trained on four of these folds and validated on the remaining one. This process is repeated five times, with each fold serving as the validation set once. This approach ensures that every data point is used both for training and validation, providing a comprehensive evaluation of the model's performance. The repeated training and validation cycles allow for a robust assessment of how well the model generalizes, highlighting any potential overfitting or underfitting issues. During each fold of cross-validation, the entire preprocessing pipeline, including imputation, scaling, and encoding, is applied to the training data. This ensures that the validation data remains unseen during the preprocessing phase, preventing data leakage and providing a more accurate evaluation of the model's performance.

The primary evaluation metric used in this cross-validation process is accuracy, which measures the proportion of correctly classified instances. However, given the imbalanced nature of the dataset, additional metrics such as precision, recall, and F1-score are also calculated to provide a more nuanced understanding of the model's performance. These metrics are crucial in fraud detection, where false positives and false negatives have different implications and costs. Finally, the results from the cross-validation are aggregated to provide an overall estimate of the model's performance. The mean accuracy, along with the standard deviation, is reported to give an indication of the model's consistency across different folds. By thoroughly validating the model through this rigorous experimental setup, the study ensures that the conclusions drawn are reliable and applicable to real-world scenarios.

2.5 Evaluation Metrics

In the context of fraud detection, choosing the right evaluation metrics is crucial due to the inherent challenges posed by imbalanced datasets. Fraudulent transactions typically constitute a small fraction of the total dataset,

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

making it essential to use metrics that not only capture the overall performance but also reflect the model's ability to correctly identify these rare but critical instances. The evaluation of the models in this study involves an accuracy as the most straightforward metric, defined as the ratio of correctly classified instances to the total number of instances. It is often the first metric considered when evaluating a model. Accuracy gives a general sense of how well the model is performing overall.

2.6 Model Implementation

The implementation of the hybrid model, alongside the comparison with traditional models, follows a systematic approach leveraging Python and its extensive ecosystem of data science libraries. The key to a successful implementation lies in the seamless integration of data preprocessing, model training, and evaluation within a cohesive pipeline, ensuring that the process is both efficient and reproducible. The entire workflow begins with data preprocessing, where the dataset undergoes a series of transformations as described earlier, including imputation of missing values, scaling of numerical features, and encoding of categorical variables. These steps are crucial for preparing the data in a format that is suitable for machine learning algorithms. To manage these preprocessing steps efficiently, the scikit-learn library's Pipeline and ColumnTransformer utilities are employed. These tools allow for the creation of a robust preprocessing pipeline that can be consistently applied across different datasets and models, minimizing the risk of data leakage, a common pitfall where information from the validation set inadvertently influences the training process. Once the data is preprocessed, the focus shifts to the implementation of the StackingClassifier. This ensemble model is constructed by combining the predictions of several base models using a meta-model, typically Logistic Regression in this study. The base models, which include RandomForestClassifier, GradientBoostingClassifier, SVC, LGBMClassifier, and XGBClassifier, are trained independently on the training data. Their predictions are then passed to the meta-model, which learns to make the final prediction based on the outputs of the base models. The meta-model effectively synthesizes the different perspectives provided by the base models, making the ensemble more robust and accurate.

To ensure that the model generalizes well to new data, a StratifiedKFold cross-validation is applied. This method divides the data into multiple folds, training the model on different subsets and validating it on the remaining data. This approach not only provides a reliable estimate of model performance but also helps in identifying any issues related to overfitting or underfitting. During each fold of cross-validation, the entire pipeline, from data preprocessing to model training and prediction, is executed, ensuring that the evaluation is as realistic as possible. The implementation also involves careful tuning of hyperparameters for both the base models and the meta-model. Hyperparameter tuning is performed using grid search or random search methods, which explore different combinations of parameters to find the optimal settings that maximize model performance. This step is critical in enhancing the robustness of the StackingClassifier, as the choice of hyperparameters can significantly impact the model's ability to learn from the data. Finally, the performance of the hybrid model is compared with that of each base model using the evaluation metrics discussed earlier. This comparison provides insights into the advantages of using a StackingClassifier over traditional models, particularly in the context of fraud detection. By systematically implementing the models and rigorously evaluating their performance, the study ensures that the findings are both reliable and applicable to real-world scenarios. The use of Python and its libraries, such as scikit-learn, pandas, and numpy, facilitates a smooth and efficient workflow, allowing for easy replication and validation of the results.

3. RESULT AND DISCUSSION

3.1 Results

As presented in the table 1, The performance of the proposed hybrid model, along with various traditional machine learning models, was evaluated using a five-fold cross-validation approach. The primary metric for comparison was accuracy, with results expressed as the mean accuracy across the folds and the associated standard deviation. The models tested include Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, k-Nearest Neighbors (k-NN), Naive Bayes, AdaBoost, LightGBM, XGBoost, Multi-Layer Perceptron (MLP) Classifier, and the proposed Hybrid Model. The results indicate that the proposed Hybrid Model significantly outperforms the traditional models in terms of accuracy. The Hybrid Model achieves an accuracy of 0.9999 with a standard deviation of 0.0001, suggesting its superior ability to distinguish between fraudulent and non-fraudulent transactions.

Logistic Regression, often used as a baseline model in binary classification tasks, performed with high accuracy at 0.9958 ± 0.0001. While this result is strong, it highlights the limitations of simpler linear models in capturing complex patterns within the data, especially when compared to more sophisticated models like XGBoost and the proposed hybrid model. Support Vector Machine (SVM) and k-Nearest Neighbors (k-NN) both yielded strong results, with accuracy scores of 0.9961 ± 0.0000 and 0.9962 ± 0.0000, respectively. These models are known for their effectiveness in high-dimensional spaces (SVM) and their simplicity and interpretability (k-NN). However, their performance, while solid, does not match that of the ensemble methods, indicating that they may not fully capture the complexity of the data. Random Forest achieved an accuracy of 0.9982 ± 0.0001, reflecting

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

its robustness and ability to handle large datasets with multiple features. The high accuracy of Random Forest demonstrates the effectiveness of ensemble methods in reducing overfitting and improving generalization. Nevertheless, the Hybrid Model surpasses Random Forest, suggesting that the combination of multiple models can capture even more intricate patterns in the data.

Gradient Boosting and AdaBoost performed similarly, with accuracies of 0.9969 ± 0.0004 and 0.9961 ± 0.0001, respectively. Both models are iterative, learning from the mistakes of previous models, which generally enhances performance. However, the slight advantage of Gradient Boosting over AdaBoost indicates that more advanced boosting techniques can lead to better outcomes. Naive Bayes, which assumes feature independence, lagged behind the other models with an accuracy of 0.9922 ± 0.0006. While Naive Bayes can be effective in certain scenarios, its simplistic assumptions limit its effectiveness in more nuanced classification tasks like fraud detection.

LightGBM and XGBoost both demonstrated excellent performance, with XGBoost leading with an accuracy of 0.9990 ± 0.0001. These gradient boosting models are optimized for speed and accuracy, making them popular choices for high-performance machine learning tasks. The slight edge of XGBoost over LightGBM may be attributed to its more comprehensive set of features and optimizations, which make it slightly better suited for this particular dataset. The Multi-Layer Perceptron (MLP) Classifier achieved a strong accuracy of 0.9962 ± 0.0000, indicating the power of neural networks in classification tasks. However, like the other traditional models, the MLP did not outperform the Hybrid Model, highlighting the advantages of combining multiple model types to leverage their collective strengths. The superior performance of the Hybrid Model can be attributed to the ensemble approach, which combines the strengths of various base models to create a more powerful and generalizable classifier. By leveraging the complementary capabilities of models like Random Forest, Gradient Boosting, SVM, and others, the Hybrid Model can capture a broader range of patterns and relationships within the data. This diversity allows it to perform well across different scenarios, including those that may confound individual models.

The near-perfect accuracy of the Hybrid Model also suggests that it is highly effective at minimizing both false positives and false negatives, a critical requirement in fraud detection where errors can have significant financial and reputational consequences. The small standard deviation (± 0.0001) further indicates that the Hybrid Model's performance is consistent across different folds of the cross-validation, reinforcing its reliability. In contrast, while traditional models like Random Forest and XGBoost performed admirably, their inability to reach the accuracy level of the Hybrid Model underscores the limitations of relying on a single model. Even the high-performing XGBoost model, with an accuracy of 0.9990 ± 0.0001, falls slightly short, illustrating the value of model stacking in achieving incremental gains in accuracy. The results also highlight the importance of model selection and the benefits of advanced ensemble techniques in tackling complex classification tasks like fraud detection. The findings suggest that organizations seeking to implement fraud detection systems could significantly benefit from adopting hybrid models, as they offer enhanced accuracy and robustness compared to traditional approaches.

Table 1. Existing and Proposed Models Accuracy

Model	Accuracy (Mean ± Std Dev)
Logistic Regression	0.9958 ± 0.0001
SVM	0.9961 ± 0.0000
Random Forest	0.9982 ± 0.0001
Gradient Boosting	0.9969 ± 0.0004
k-NN	0.9962 ± 0.0000
Naive Bayes	0.9922 ± 0.0006
AdaBoost	0.9961 ± 0.0001
LightGBM	0.9954 ± 0.0007
XGBoost	0.9990 ± 0.0001
MLP Classifier	0.9962 ± 0.0000
Hybrid Model (Proposed)	0.9999 ± 0.0001

3.2 Trade-off Analysis

When evaluating machine learning models, particularly in the context of critical applications like fraud detection, it's essential to consider the trade-offs between different performance metrics, model complexity, computational cost, and the practical implications of false positives and false negatives. The results from the cross-validation accuracy analysis reveal several important trade-offs that must be considered when selecting the most appropriate model for deployment. The proposed Hybrid Model achieves the highest accuracy at 0.9999 ± 0.0001, significantly outperforming all individual models. However, this gain in accuracy comes with increased model complexity. The Hybrid Model combines multiple base models, each with its unique strengths, into a meta-model that aggregates their predictions. While this approach maximizes accuracy, it also introduces additional layers of complexity. This complexity can lead to longer training times, increased computational costs, and potentially greater challenges in

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

model interpretability and deployment. In contrast, simpler models like Logistic Regression, with an accuracy of 0.9958 ± 0.0001, offer easier interpretability and faster training times but at the cost of slightly lower performance.

The models differ in their computational efficiency, which impacts their suitability for real-time fraud detection systems. Simpler models such as Logistic Regression and Naive Bayes are computationally efficient, making them suitable for environments where resources are limited or quick decision-making is crucial. These models can process transactions rapidly, but they may not capture the full complexity of the data, leading to lower overall accuracy. On the other hand, models like XGBoost and the Hybrid Model, while offering superior performance, require more computational power and time. XGBoost, with its accuracy of 0.9990 ± 0.0001, is known for its speed and efficiency relative to other gradient boosting models, yet it still demands significant computational resources. The Hybrid Model, being the most complex, combines several algorithms, increasing the computational burden even further. This increased demand may be justified in scenarios where the highest accuracy is critical, but it could be a limiting factor in environments where computational resources or time are constrained.

The interpretability of a model is an essential factor, especially in domains like finance, where understanding the decision-making process is crucial for compliance and trust. Simpler models like Logistic Regression and even Decision Trees within a Random Forest offer greater interpretability, allowing stakeholders to understand and explain why certain transactions are flagged as fraudulent. These models, however, sacrifice some predictive power compared to more complex models. The Hybrid Model, with its layered structure of base models and a meta-model, offers the highest predictive power but at the cost of reduced interpretability. Understanding the decision-making process in such a complex model is challenging, as the final prediction is a result of multiple interacting models. This lack of transparency can be a drawback in situations where model decisions need to be audited or explained to non-technical stakeholders.

3.3 Constraints and Limitations

The proposed Hybrid Model has demonstrated superior performance in detecting fraudulent transactions, but several constraints and limitations must be acknowledged. These limitations are crucial for interpreting the results of this study and considering the practical application of the model in real-world scenarios. One of the primary limitations is the high computational cost associated with the Hybrid Model. Its complexity, which involves training multiple base models and a meta-model, demands significant computational power and time. This poses a challenge, especially for organizations with limited resources. The need for extensive cross-validation to ensure model robustness further increases the computational demands, potentially making it difficult to deploy such a model in real-time environments without substantial investment in computational infrastructure.

Another significant limitation is the reduced interpretability of the Hybrid Model. Due to its ensemble nature, combining several different algorithms, the decision-making process becomes more opaque. While techniques like SHAP (SHapley Additive exPlanations) can provide some level of interpretability, they add complexity and may not fully satisfy the transparency requirements in industries where explainability is crucial. In finance, for example, where regulatory compliance often demands clear and interpretable outputs, the lack of transparency in the Hybrid Model could limit its acceptance and usability. The performance of the Hybrid Model is also highly dependent on the quality and nature of the data used for training. While the model performed exceptionally well on the dataset used in this study, its effectiveness may vary with different datasets, particularly those with different distributions or feature types. The model's reliance on specific preprocessing steps tailored to this dataset limits its generalizability, meaning that if applied to data with different characteristics, such as from a different financial institution or geographic region, additional tuning and validation would be required to maintain performance.

Another limitation arises from the inherent imbalance in the dataset, where fraudulent transactions are significantly outnumbered by non-fraudulent ones. Although the Hybrid Model is designed to handle imbalanced data through ensemble learning techniques, there remains a risk of bias towards the majority class. While the model's high accuracy suggests effective handling of this imbalance, accuracy alone may not fully capture the model's performance in detecting rare fraudulent cases. Metrics such as precision, recall, and F1-score offer a more nuanced view, but the challenge of imbalanced data remains a critical constraint.

Given the complexity of the Hybrid Model, there is also an inherent risk of overfitting, particularly if the model is not carefully tuned. Overfitting occurs when the model learns to perform exceptionally well on the training data but fails to generalize to unseen data. Although cross-validation was employed to mitigate this risk, overfitting could still pose a problem, especially if the model encounters a significantly different data environment in a real-world application. This underscores the importance of continuous monitoring and validation when deploying the model in practice. Scalability is another important consideration. While the Hybrid Model has shown outstanding performance, its complexity could limit its scalability, particularly in environments requiring real-time processing of large data volumes. The computational demands and time required to train and update the model could be prohibitive, especially for applications that need to scale rapidly in response to growing data volumes. In such cases, simpler models or optimized versions of ensemble methods may be more appropriate. Lastly, the generalization of the Hybrid Model to domains beyond fraud detection remains uncertain. The model was specifically designed and tested for financial transaction data, and its effectiveness in other types of classification

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

tasks, such as medical diagnosis or customer segmentation, has not been tested. Applying the Hybrid Model to different domains may require significant adjustments in data preprocessing, feature selection, and model tuning, potentially limiting its broad applicability.

4. CONCLUSION

This study explored the development and evaluation of a Hybrid Model for fraud detection, demonstrating that the proposed approach significantly outperforms traditional machine learning models in terms of accuracy, achieving a near-perfect classification accuracy of 99.99%. By combining multiple base models such as Random Forest, Gradient Boosting, SVM, LightGBM, and XGBoost through a stacking technique, the Hybrid Model effectively leverages the strengths of each algorithm. The findings suggest that ensemble methods, particularly hybrid approaches, offer substantial advantages in handling the complexities and nuances of fraud detection tasks. However, the superior performance of the Hybrid Model comes with trade-offs, including increased computational demands and reduced interpretability, which are significant considerations, especially for organizations with limited resources or strict regulatory requirements. The model's dependency on the specific characteristics of the training data, as well as challenges posed by imbalanced datasets, highlights the importance of careful tuning and validation. Despite these limitations, the Hybrid Model presents a powerful tool for enhancing fraud detection systems, offering a level of accuracy that could significantly reduce financial losses and improve security in financial transactions. Future research could explore ways to optimize the Hybrid Model further, potentially reducing its computational overhead and improving its interpretability without compromising performance. In conclusion, the Hybrid Model represents a significant advancement in fraud detection, offering a robust and effective solution for identifying fraudulent transactions with a potential accuracy improvement of up to 1% over traditional models. Its success demonstrates the potential of hybrid approaches in machine learning, paving the way for more sophisticated and reliable models in the field of fraud detection and beyond, where thoughtful consideration of its limitations and careful implementation could play a crucial role in safeguarding financial systems from the ever-evolving threat of fraud.

REFERENCES

[1] J. Nicholls, A. Kuppa, and N.-A. Le-Khac, “Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape,” *Ieee Access*, vol. 9, pp. 163965–163986, 2021.
[2] O. A. Bello, A. Ogundipe, D. Mohammed, F. Adebola, O. A. Alonge, and others, “AI-Driven approaches for real-time fraud detection in US financial transactions: challenges and opportunities,” *Eur. J. Comput. Sci. Inf. Technol.*, vol. 11, no. 6, pp. 84–102, 2023.
[3] P. Chatterjee, D. Das, and D. B. Rawat, “Digital twin for credit card fraud detection: Opportunities, challenges, and fraud detection advancements,” *Futur. Gener. Comput. Syst.*, 2024.
[4] I. H. Sarker, “Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions,” *SN Comput. Sci.*, vol. 2, no. 6, p. 420, 2021.
[5] M. S. Ibrahim, W. Dong, and Q. Yang, “Machine learning driven smart electric power systems: Current trends and new perspectives,” *Appl. Energy*, vol. 272, p. 115237, 2020.
[6] J. Sansana *et al.*, “Recent trends on hybrid modeling for Industry 4.0,” *Comput. \& Chem. Eng.*, vol. 151, p. 107365, 2021.
[7] M. Seera, C. P. Lim, A. Kumar, L. Dhamotharan, and K. H. Tan, “An intelligent payment card fraud detection system,” *Ann. Oper. Res.*, vol. 334, no. 1, pp. 445–467, 2024.
[8] R. Thakur and D. Rane, “Machine learning and deep learning for intelligent and smart applications,” in *Future Trends in 5G and 6G*, CRC Press, 2021, pp. 95–113.
[9] X. Larriva-Novo, M. Vega-Barbas, V. A. Villagra, D. Rivera, M. Alvarez-Campana, and J. Berrocal, “Efficient distributed preprocessing model for machine learning-based anomaly detection over large-scale cybersecurity datasets,” *Appl. Sci.*, vol. 10, no. 10, p. 3430, 2020.
[10] A. Aljohani, “Predictive analytics and machine learning for real-time supply chain risk mitigation and agility,” *Sustainability*, vol. 15, no. 20, p. 15088, 2023.
[11] M. Sánchez-Aguayo, L. Urquiza-Aguiar, and J. Estrada-Jiménez, “Fraud detection using the fraud triangle theory and data mining techniques: A literature review,” *Computers*, vol. 10, no. 10, p. 121, 2021.
[12] T. Kavzoglu and A. Teke, “Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost),” *Arab. J. Sci. Eng.*, vol. 47, no. 6, pp. 7367–7385, 2022.
[13] O. San, A. Rasheed, and T. Kvamsdal, “Hybrid analysis and modeling, eclecticism, and multifidelity computing toward digital twin revolution,” *GAMM-Mitteilungen*, vol. 44, no. 2, p. e202100007, 2021.
[14] F. Guo *et al.*, “A Hybrid Stacking Model for Enhanced Short-Term Load Forecasting,” *Electronics*, vol. 13, no. 14, p. 2719, 2024.
[15] V. F. Rodrigues *et al.*, “Fraud detection and prevention in e-commerce: A systematic literature review,” *Electron. Commer. Res. Appl.*, vol. 56, p. 101207, 2022.
[16] K. Khando, M. S. Islam, and S. Gao, “The emerging technologies of digital payments and associated challenges: a systematic literature review,” *Futur. Internet*, vol. 15, no. 1, p. 21, 2022.
[17] M. M. Alam, A. E. Awawdeh, and A. I. Bin Muhamad, “Using e-wallet for business process development: challenges and prospects in Malaysia,” *Bus. Process Manag. J.*, vol. 27, no. 4, pp. 1142–1162, 2021.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1118-1127
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5840

[18] S. O. Pinto and V. A. Sobreiro, “Literature review: Anomaly detection approaches on digital business financial systems,” *Digit. Bus.*, vol. 2, no. 2, p. 100038, 2022.
[19] A. K. Mishra, S. Anand, N. C. Debnath, P. Pokhariyal, and A. Patel, *Artificial Intelligence for Risk Mitigation in the Financial Industry*. John Wiley & Sons, 2024.
[20] M. McLennan and others, “The global risks report 2022 17th edition,” 2022.
[21] S. Abimannan, E.-S. M. El-Alfy, Y.-S. Chang, S. Hussain, S. Shukla, and D. Satheesh, “Ensemble multifeatured deep learning models and applications: A survey,” *IEEE Access*, 2023.
[22] G. Kunapuli, *Ensemble methods for machine learning*. Simon and Schuster, 2023.
[23] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, “Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning,” *Sustainability*, vol. 14, no. 21, p. 13998, 2022.
[24] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Credit Card Fraud Detection," Kaggle, 2015. [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. [Accessed: Aug. 23, 2024].