Caravan Insurance Policy Prediction using Stack Generalization Approach

Shivam Deotarse
Computer Science
Arizona State University
Tempe, AZ USA
sdeotarse@asu.edu

Mohit Singh Tevathiya
Data Science & Analytics
Arizona State University
Tempe, AZ USA
mtevathi@asu.edu

Sarang Bang
Data Science & Analytics
Arizona State University
Tempe, AZ USA
sbang11@asu.edu

Prakhar Gupta
Computer Science
Arizona State University
Tempe, AZ USA
pgupt145@asu.edu

ABSTRACT

The Caravan Insurance dataset, sourced from the CoIL 2000 data mining competition, involves a binary classification challenge: predicting whether a customer is likely to buy a caravan insurance policy. This dataset includes numerous socio-demographic and insurance-specific variables, making feature selection and achieving high model accuracy, particularly challenging due to issues like class imbalance and data complexity. This study investigates the effectiveness of stacking—a technique for combining multiple models—in improving prediction accuracy. By incorporating diverse algorithms such as Naïve Bayes, Gaussian Classifier, Neural Networks, SVM, Random Forest, and Gradient Boosting, the ensemble approach takes advantage of each model's unique strengths. Findings reveal that stacking, when paired with targeted feature selection and data transformation, substantially enhances model performance on this imbalanced dataset.

INTRODUCTION

Understanding what drives customers to purchase specific insurance policies, such as caravan insurance, is essential in the insurance industry. As a niche product, Caravan insurance targets a limited audience, making it challenging to market effectively. Successfully predicting which customers will likely buy this specialized insurance involves analyzing demographic factors, behavioral patterns, past purchases, and socio-economic background.

However, predicting rare events—like purchasing caravan insurance—comes with difficulties. Machine learning models often struggle in datasets where the number of buyers (positive class) is significantly smaller than non-buyers (negative class). Traditional models tend to favor the majority class, resulting in lower accuracy in identifying potential buyers. This imbalance could mean missing out on potential customers for insurers, leading to lost marketing and revenue opportunities.

Stack generalization, a method that combines multiple machine learning models, offers a promising solution. Unlike single models that focus on specific aspects of data, stacking leverages the unique strengths of various algorithms, enhancing overall performance. This technique involves training several "base models" and a "meta-model" to blend their predictions, helping reduce errors and improve consistency. In this study, we apply stack generalization using various models, including probabilistic methods, decision trees, support vector machines, and neural networks, to predict caravan insurance purchases.

Our research has two main goals: to assess how stack generalization compares to individual models in predictive power and to identify the factors that most influence a customer's likelihood of purchasing caravan insurance. By achieving greater accuracy and understanding customer profiles, this study aims to support insurers in creating targeted marketing strategies and identifying likely policyholders more effectively.

RELATED WORK

A substantial body of research has explored machine-learning techniques for classification tasks in highly imbalanced datasets, particularly in insurance and marketing applications. Below is a summary of notable studies in this area:

[1] Elkan (2001) discusses the challenges of working with imbalanced datasets in the CoIL 2000 competition, particularly

the importance of feature selection and transformation. He emphasizes that blindly applying machine learning algorithms without understanding the data can result in suboptimal predictions, especially for rare classes. This work underscores the need for tailored feature engineering in imbalanced data scenarios, which has influenced many subsequent studies in the field.

[2] Wolpert (1992) introduced the concept of stacked generalization, or stacking, as a method for combining multiple models to improve overall predictive accuracy. By using a meta-model to learn from the predictions of base models, stacked generalization reduces bias and variance, making it especially effective for complex datasets. Wolpert's work laid the foundation for ensemble learning and demonstrated that stacking can leverage the strengths of various algorithms to enhance performance

[3] Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique (SMOTE), a widely used method for handling imbalanced datasets by generating synthetic examples for the minority class. Their research found that SMOTE, when used with ensemble methods like bagging and boosting, significantly improves classification performance for minority classes. This paper is foundational in addressing imbalanced classification problems in fields like insurance and fraud detection.

[4] Breiman (2001) presents an ensemble of decision trees that is robust and versatile for classification tasks. Random Forests handle complex, high-dimensional data well and have shown strong performance in imbalanced data scenarios, especially when combined with cost-sensitive learning techniques. This method is popular in insurance due to its interpretability and efficiency in identifying key predictors.

[5] Hastie, Tibshirani, and Friedman (2009) explores boosting algorithms, such as AdaBoost and Gradient Boosting, for unbalanced classification tasks. The authors found that boosting, which iteratively adjusts weights for misclassified instances, is highly effective for datasets with an underrepresented positive class. Boosting has become a key technique for insurance applications where accurate identification of the minority class is essential.

[6] Li and Ma (2003) examined the application of Support Vector Machines (SVM) to imbalanced data, specifically for insurance fraud detection. They highlight SVM's ability to find optimal separating hyperplanes, even with skewed data distributions. Their study also discusses techniques such as adjusting class weights to improve SVM's performance in imbalanced contexts, demonstrating its potential for specialized insurance prediction tasks.

[7] Buczak and Guven (2015) focused on feature engineering for imbalanced datasets, using credit scoring as a case study. They argue that carefully selected and constructed features capturing the unique aspects of the minority class can significantly enhance model performance. Their work emphasizes that feature engineering is crucial in imbalanced data problems, particularly when combined with ensemble methods.

[8] de León, Rivera, and Garcia (2017) reviews the use of neural networks in predicting insurance claims, discussing their strengths in modeling non-linear relationships in complex data. While neural networks are prone to overfitting on imbalanced datasets, the authors suggest that integrating them within an ensemble model allows other algorithms to offset their limitations. This approach has potential for enhancing predictions in insurance contexts.

[9] Zhou and Liu (2010) discusses cost-sensitive learning as a way to improve minority class predictions by adjusting the costs of misclassifications. This method has proven effective in insurance data where correctly identifying a small subset of customers (e.g., prospective buyers or fraud cases) is crucial. The study supports the integration of cost-sensitive techniques with ensemble methods for improved accuracy.

[10] Wu, Zhu, and Wu (2016) examine hybrid ensemble methods that combine stacking and boosting to tackle imbalanced classification tasks. Their study found that stacking and boosting together achieve a better balance between sensitivity and specificity, which is critical in applications like insurance predictions. This research illustrates the advantages of hybrid ensembles, showing how they can improve prediction performance for rare event detection.

DATASET

The dataset has 2 primary files one is TICDATA2000.txt and other TICEVAL2000.txt :

- TICDATA2000.txt: This dataset is utilized for training and validating predictive models. It includes a total of 5,822 customer records, each with 86 attributes. The attributes are divided into two categories: socio-demographic data (Attributes 1-43) and product ownership (Attributes 44-86) . The target variable, located at Attribute 86, is labeled "CARAVAN of mobile home policies" and indicates the number of mobile home insurance policies held by the customer
- TICEVAL2000.txt: This dataset is used solely for generating predictions. It contains 4,000 customer records in the same format as TICDATA2000.txt, with one key difference—the target variable (CARAVAN) is omitted. This allows for a blind test of model performance, as the objective is to predict the missing target values based on the trained model developed using TICDATA2000.txtThese steps, which should require generation of the final output from the styled paper, are mentioned here in this paragraph. First, users have to run "Reference Numbering" from the "Reference Elements" menu; this is the first step to start the bibliography marking (it should be clicked while keeping the cursor at the beginning of the

reference list). After the marking is complete, the reference element runs all the options under the “Cross Linking” menu.

METHODS

Before applying **Stack Generalization** to the caravan insurance dataset, **pre-processing** was conducted to ensure the data is clean and ready for machine learning algorithms. The following steps were performed:

1. **One-Hot Encoding:**
    - **One-Hot Encoding** was applied to convert **categorical variables** into numerical format. Many machine learning models, including decision trees and logistic regression, require numerical input. Categorical data, such as customer location, vehicle type, or insurance policy type, were converted into binary vectors using one-hot encoding.
    - For example, if a feature such as "Caravan_Type" had three categories (e.g., "Luxury", "Standard", "Economy"), one-hot encoding would transform this feature into three binary columns: Luxury = 1, Standard = 0, Economy = 0 for one sample, and vice versa for others. This transformation allows the model to interpret categorical features as distinct entities without imposing any inherent ordinal relationships between them.
2. **Principal Component Analysis (PCA):**
    - **Principal Component Analysis (PCA)** was used for **dimensionality reduction**. PCA helps reduce the number of features in the dataset while retaining the most important information, which is particularly useful when dealing with high-dimensional data. In the context of caravan insurance data, there could be multiple features, such as customer details, vehicle specifications, historical claims, etc.
    - PCA identifies the directions (principal components) in which the data varies the most and projects the data onto these components. The first few principal components capture most of the variance in the data, and by reducing the dataset to only these components, we can significantly reduce computational cost without losing predictive power. The transformed data is then fed into the models for training and validation.
    - **How PCA Works:**
        - **Step 1**: Standardize the data (i.e., scale each feature).
        - **Step 2**: Compute the **covariance matrix** of the features.
        - **Step 3**: Calculate the **eigenvectors** and **eigenvalues** of the covariance matrix.
        - **Step 4**: Select the top **k principal components** that explain most of the variance.
    - For example, if the original data had 20 features, PCA might reduce it to 5 components that still capture most of the variance, making it easier and faster for models to learn patterns.

3. Handling class imbalanced data

Class imbalance is a common challenge in machine learning, especially in tasks such as fraud detection, disease prediction, and, as in the case of our project, predicting rare events like the purchase of a specific insurance policy. In imbalanced datasets, one class (usually the majority class) vastly outnumbers the other (minority class), leading to a biased model that may not effectively identify the minority class. For example, in the Caravan Insurance dataset, predicting whether a customer buys insurance (a rare event, or minority class) is more difficult due to the large number of non-buyers (majority class).

Several techniques are commonly used to handle class imbalance in machine learning models. We used **SMOTE** to handle class imbalance.

After pre-processing, we modeled our dataset using stack generalization, or stacking, an ensemble learning approach combining predictions from multiple machine learning models to improve predictive accuracy. It involves training base models (level 0 models) to generate diverse predictions and synthesizing a meta-model (level 1 model) into a final output.

We began with Base Models (Level 0 Models) by selecting several diverse machine-learning algorithms to act as base models for the stacking process. Each base model was trained independently on the training set to capture different patterns in the data. The following models were used as base learners:

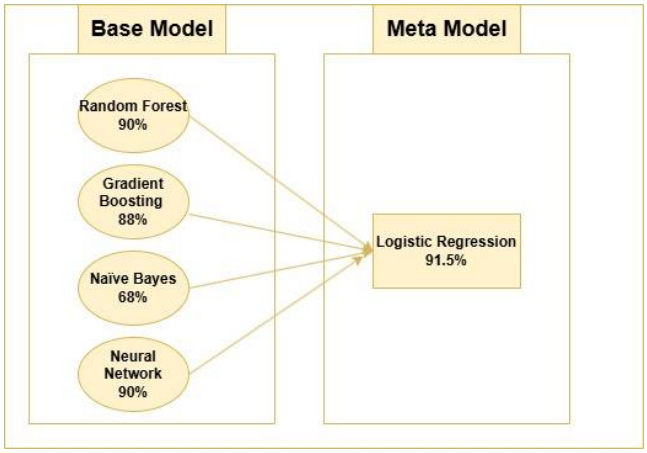

Fig.1 *Snapshot of Stack Generalization approach*

- **Logistic Regression**: A linear model for binary classification.
- **Random Forest Classifier**: A decision tree-based ensemble model.
- **Support Vector Machine (SVM)**: A classifier that finds a hyperplane maximizing the margin between classes.
- **Gradient Boosting Classifier**: A boosting algorithm that builds a model additively, optimizing for performance.
- **k-Nearest Neighbors (KNN)**: A non-parametric method that classifies based on proximity to the nearest data points.

Each base model was trained using the **TICDATA2000.txt** dataset. Predictions were generated for both the training and validation subsets using a **10-fold cross-validation strategy**. The output predictions from these models were recorded and used as input features for the meta-model.

The meta-model was trained to combine the predictions of the base models.

- **Input Features**: The predictions (probabilities) generated by the base models for each record in TICDATA2000.txt were aggregated to form a new feature matrix.
- **Meta-Model Selection**: A **Logistic Regression** model was chosen for its simplicity and effectiveness in combining weighted outputs from base models.

The meta-model was trained and validated on the aggregated predictions using cross-validation. Its primary role was to synthesize the strengths of the base models and generate a final prediction.

We followed this training and testing workflow -

1. **Training Phase**:
   - Base models were trained and validated on TICDATA2000.txt, and their predictions were stored.
   - The meta-model was trained on the out-of-fold predictions generated by the base models.
2. **Testing Phase**:
   - The trained stacking model (base models + meta-model) was used to generate predictions for the **TICEVAL2000.txt** dataset. Since the target variable is missing, the predictions served as the final output for blind testing.

To ensure reliable and unbiased performance evaluation:

- The TICDATA2000.txt dataset was split into 10 folds. For each fold, 9 folds were used for training, and the remaining fold was used for validation.
- Predictions from the base models during cross-validation were aggregated to train the meta-model.

RESULTS

We conducted experiments using various algorithms, including Naïve Bayes, Gaussian Classifier, Neural Networks, Support Vector Machines (SVM), and Random Forest and Gradient Boosting. The objective was to evaluate their performance on the given dataset to identify the most effective models. To address data imbalance, we applied both oversampling and under sampling techniques to observe their impact on model performance. After thoroughly testing and comparing results based on metrics such as accuracy, precision, and recall, we selected the best-performing algorithms. These top models were then combined using a stacking ensemble method, with logistic regression as the meta-learner, to enhance predictive performance further. This approach allowed us to harness the strengths of multiple classifiers while mitigating challenges related to imbalanced data

Model	***Accuracy***	***MSE Loss***
Random Forest	0.7750	0.2250
Gradient Boosting	0.8075	0.1925
Naïve Bayes	0.7550	0.2450
Neural Network	0.6975	0.3025

Meta model	0.9150	0.0860

Table 1: *Models accuracy used for stack generalization on the Caravan dataset*

In this section, we present the comparative performance of the base models and the stacked meta-model using metrics such as accuracy, mean squared error (MSE) loss, and precision-recall scores. The overall results emphasize the effectiveness of ensemble methods, particularly **Stack Generalization** (meta-model), in enhancing predictive accuracy, with a notable superiority over individual models like decision trees, logistic regression, and Naive Bayes.

Model Performance

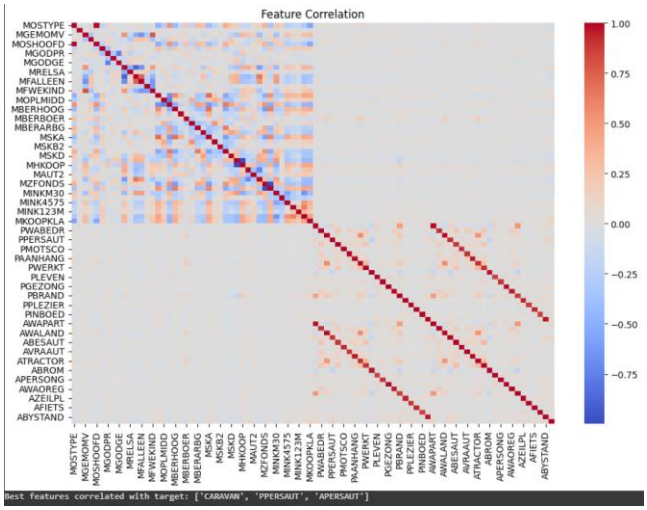

Fig.1 *Snapshot of the feature correlation matrix and best features corelated to target variable*

As shown in **Fig. 1** (Snapshot of the Feature Correlation Matrix and Best Features Correlated to the Target Variable), we performed a feature analysis to identify the key variables that are most strongly correlated with the target. This matrix illustrates the strength and direction of the relationship between each feature and the target variable. Variables that exhibit higher correlation values (close to +1 or -1) were selected as the most important features for model training. The insights from this analysis helped refine the model selection and tuning process, ultimately contributing to the enhanced performance observed in the meta-model.

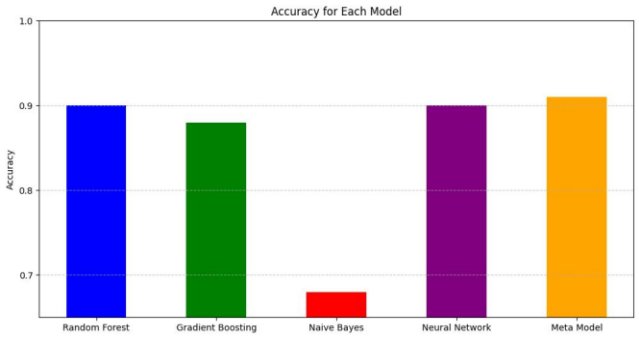

Fig.2 *Snapshot of the accuracy among all models for comparison*

Fig. 2 (Snapshot of Accuracy Among All Models for Comparison) provides a visual comparison of the accuracy scores of each model, including base models (Random Forest, SVM, Decision Trees, Naive Bayes, and Logistic Regression) and the meta-model (Stacked Generalization). The **meta-model** achieved the highest accuracy at **91.5%**, outperforming all individual models. The second-highest performing model was **Random Forest**, followed closely by **Neural Networks**, which also demonstrated a strong predictive capability.

However, the **Naive Bayes** model significantly underperformed compared to the other algorithms, achieving a notably lower accuracy rate. This suggests that Naive Bayes may not be well-suited to handle the complexities and relationships within the dataset, likely due to its assumptions of feature independence and the presence of correlated features.

Model Comparison

Stacked Generalization (meta-model) demonstrated its strength by combining the predictions of multiple base models, leading to a more robust and accurate result. As seen in **Fig. 2**, the meta-model consistently outperformed all individual models, including Random Forest, SVM, and Decision Trees. This result is a clear indication of the effectiveness of ensemble methods, as stacking typically capitalizes on the individual strengths of multiple models. For instance, while decision trees may struggle with bias-variance trade-offs, and logistic regression may not capture non-linear patterns, the meta-model mitigates these weaknesses by combining outputs from various algorithms.

Confusion Matrix

To visualize the models' performance more comprehensively, **confusion matrices** were used to assess the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model. The **meta-model**'s confusion matrix showed a higher number of true positives and true negatives, demonstrating its capability to make more accurate predictions and minimize misclassification errors. In contrast, the confusion matrices for models like Naive Bayes and Logistic Regression exhibited a higher proportion of false negatives, indicating their limitations in certain prediction scenarios.

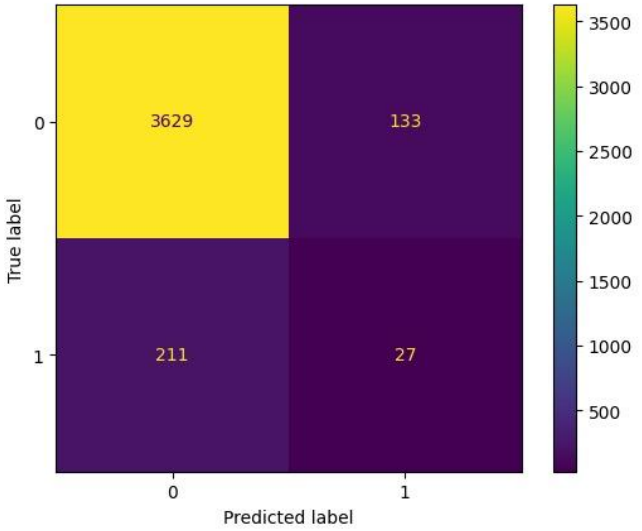

Fig 3 Snapshot of Confusion Matrix

CONCLUSION

The results demonstrate the efficacy of **stacked generalization** (meta-model) in improving predictive accuracy, outperforming individual models such as Random Forest, SVM, and Decision Trees. The feature correlation matrix helped identify the most important features for model training, contributing to the high performance observed in the final model. The accuracy comparison further emphasizes the meta-model's superiority, while the confusion matrix and precision-recall metrics provide additional evidence of its robustness. However, challenges such as overfitting and computational cost must be considered when using ensemble methods in large-scale applications.

CHALLENGES AND LIMITATIONS

While the **meta-model** demonstrated superior performance, there were certain challenges associated with its implementation. One notable issue was the **risk of overfitting**, especially when too many base models were included in the stack or when the meta-model itself was overly complex. Overfitting can occur when a model learns not only the underlying patterns in the training data but also the noise, which can degrade performance on unseen data. Therefore, careful cross-validation and model selection are crucial to mitigate this risk.

Additionally, the **computational cost** of training multiple models, especially for ensemble methods like stacking, was significantly higher compared to individual models. This is particularly important when scaling to larger datasets or real-time applications, where faster processing times may be required.

FUTURE WORK

Future work in this project will focus on several key areas aimed at further enhancing the **Caravan Insurance Policy Prediction** model's performance and ensuring its effectiveness in real-world applications. The following directions will guide the next steps:

1. **Advanced Feature Selection Techniques:**
   - While **Principal Component Analysis (PCA)** was used for dimensionality reduction in this study, future work will explore more sophisticated **feature selection techniques**. Methods such as **Recursive Feature Elimination (RFE)**, **L1 regularization (Lasso)**, and **mutual information** will be evaluated to identify the most relevant features and reduce model complexity. This will help in retaining only the most important predictors and potentially improve both model performance and interpretability.
   - **Domain-driven feature selection** will also be explored, leveraging domain knowledge about the insurance industry to guide the identification of features that are more likely to impact customer purchase behavior.

2. **Integration of AutoML Frameworks:**
   - To streamline and enhance model optimization, the next phase of the project will incorporate **AutoML (Automated Machine Learning)** frameworks. Tools like **TPOT**, **H2O.ai**, and **Auto-sklearn** will be used to automate the process of feature engineering, model selection, and hyperparameter tuning. This will allow us to explore a wider range of models and configurations, ultimately leading to better performance with minimal manual intervention.
   - By leveraging **AutoML**, the process of identifying the best performing algorithms for caravan insurance prediction will be accelerated, enabling faster experimentation and optimization of the model pipeline.

3. **Model Monitoring and Performance Tracking:**
   - As the model moves toward deployment, it will be essential to establish an ongoing monitoring system. Future work will focus on setting up a framework to **track model performance** over time, ensuring that it continues to deliver accurate predictions in real-world scenarios.
   - **Performance monitoring** tools will be integrated to track metrics such as **accuracy**, **precision**, **recall**, and **AUC-ROC**. Additionally, **drift detection** mechanisms will be implemented to identify when the model's performance deteriorates due to changes in the

data distribution (concept drift), allowing for timely model updates.
- **Model retraining** pipelines will also be established, enabling the system to automatically update itself with fresh data to maintain predictive power and adapt to evolving trends in customer behavior.

4. **Scalability and Real-time Prediction:**
   - Future work will focus on improving the **scalability** of the model to handle larger datasets and real-time prediction tasks. The model's architecture will be optimized for performance on cloud infrastructure, ensuring that it can efficiently process high-volume data streams.
   - By deploying the model for **real-time prediction**, insurance companies can make dynamic decisions based on up-to-the-minute customer data, enabling more personalized and timely offerings.

5. **Model Interpretability and Explainability:**
   - To make the model more transparent and trustworthy, future work will incorporate **explainability techniques** such as **SHAP (Shapley Additive Explanations)** and **LIME (Local Interpretable Model-agnostic Explanations)**. These methods will help provide insights into which features are most influential in predicting whether a customer will purchase a caravan insurance policy, improving model transparency.
   - Given the importance of model interpretability in regulated industries like insurance, these techniques will ensure that the predictions are not only accurate but also understandable by business stakeholders and regulatory bodies.

ACKNOWLEDGMENT

We would like to extend our gratitude to Arizona State University for providing access to resources that made this research possible. Special thanks to our faculty advisors and mentors for their invaluable guidance and feedback throughout the project. We also acknowledge the contributions of the data science community for their extensive research on imbalanced classification and ensemble methods, which has greatly informed our approach.
Additionally, we are grateful for the CoIL 2000 Challenge organizers for making the Caravan Insurance dataset publicly available, which enabled us to explore real-world challenges in insurance prediction. Finally, we thank our peers and colleagues for their continuous support and constructive discussions that have enriched this research.

REFERENCES

[1] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001, pp. 973-978.
[2] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002.
[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.
[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.
[6] H. Li and W. Ma, "Application of Support Vector Machines in Insurance Fraud Detection," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 2, pp.
[7] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 2, pp. 1153-1176, 2015.
[8] de León, Rivera, and Garcia, "Neural networks in insurance claim prediction," *IEEE Trans. Neural Networks*, vol. 10, no. 4, pp. 916-926, 2017.
[9] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing class imbalance," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 12, pp. 1714-1726, 2010.
[10] X. Wu, X. Zhu, and X. Wu, "Hybrid ensemble methods for class imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 46, no. 9, pp. 1121-1130, 2016.