International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V71I4P206

Volume 71 Issue 4, 65-71, April 2023


*Original Article*

# A Novel Multi-Stage Stacked Ensemble Classifier using Heterogeneous Base Learners

N. Pavitha[1], Shounak Sugave[2]

[1,2]*School of Computer Engineering and Technology, MIT World Peace University, Pune, Maharashtra, India*
[1]*Department of Computer Engineering, Vishwakarma University, Pune, Maharashtra, India*

[1]*Corresponding Author : pavithanrai@gmail.com*

Received: 07 November 2022 Revised: 04 February 2023 Accepted: 13 April 2023 Published: 25 April 2023

***Abstract*** *- The assessment of credit risk is essential to contemporary economies. Historically, statistical techniques and manual auditing have been used to measure credit risk associated with credit grants. Recent developments in financial AI are the result of a new generation of machine learning (ML)-driven credit risk models that have drawn a lot of interest from both business and academia. In this research, we aimed to improve ML algorithms' performance by optimizing hyperparameters. Also, we proposed a novel multi-stage heterogeneous stacked ensemble ML algorithm for predicting credit risk. Experimental results show a significant improvement in the performance of the proposed algorithm as compared to other hyperparameter-optimized ML algorithms. Two real-time data sets from emerging market economies are used to evaluate our model. Different evaluation metrics, namely precision, recall, f1-score and accuracy, are used for evaluating model performance.*

***Keywords*** *- Machine learning, Ensemble algorithm, Heterogeneous ensemble, Statistical modelling, Credit risk.*

## 1. Introduction and Literature

Most financial organizations use an individual's financial credibility to determine whether or not to grant a loan (Lin et al., 2012). A well-designed credit scoring model that can distinguish between applicants with good credit and unacceptable can reduce potential customer churn (Dutta et al., 2022) and generate enormous profits for a business. In contrast, a poorly designed model that cannot make this distinction can result in enormous financial losses.

Nowadays, a person's credit score is frequently used to assess their financial reliability. Logistic regression and linear discriminant analysis are two statistical techniques that have historically been used in credit risk approaches. (Bhattacharya et al., 2022; Markov et al., 2022) However, handling huge datasets is difficult for these strategies. With the advancement in AI, machine learning algorithms such as KNN, RF and SVM and deep learning algorithms are used to solve credit risk prediction problems, particularly when the data set is imbalanced (Lenka et al., 2018).

In the research paper (Carta et al., 2021), the authors used ensemble techniques with feature space which was discretized for the credit score calculation. Authors proposed balanced sampling (Zhang, Yang, & Zhang, 2021) and a voting-based algorithm was proposed for outlier detection. Authors claim that the outlier score has enhanced the performance. The bagging approach of the ensemble was used by (Zhang, Yang, Zhang, et al., 2021) for the credit score calculation. A genetic algorithm-based ensemble technique was proposed by (Jin et al., 2021) for improved credit risk prediction. A stacking ensemble approach for noise reduction is proposed by (Yao et al., 2022). Clustering based ensemble approach was proposed by (Singh et al., 2021) for credit risk prediction problems.

Researchers proved that ensemble approaches significantly improve the performance of algorithms in various domains. (Lakshmanarao & Shashi, 2022; Sunitha Bai & Malempati, 2022; Yamashkin et al., 2022). Authors in the research paper (Tripathi et al., 2021) also carried out a comprehensive survey on various ensemble techniques, and a review of ensemble techniques for credit scoring problems was done (Lenka et al., 2022). In this research, the stacked ensemble technique is applied to the banking and financial sector risk prediction problem. The rest of the paper is organized as follows. In the next section, section 2, various ensemble approaches are elaborated. Section 3 speaks about various data sets used for the study—section 4 covers the methodology adopted for the research. Section 5 discusses the results, and finally, section 5 concludes the research paper and future research directions.

## 2. Ensemble Approaches

Ensemble methods are techniques for improving model accuracy by combining multiple models rather than using a single model. The combined models significantly improve the accuracy of the results. Ensemble techniques are categorized as bagging, boosting and stacking techniques.

### *2.1. Bagging*

The individual model is trained separately in the bagging ensemble approach, averaging each prediction. Bagging ensembled models will have less variance than any individual model.(Sun & Pfahringer, 2011) Ultimately bagging models enhance performance and also reduce the variance.

### *2.2. Boosting*

To minimize training errors, boosting is an ensemble learning method that combines a group of weak learners into a strong learners. A random sample of data is chosen, fitted with a model, and then trained sequentially—that is, each model attempts to compensate for the shortcomings of its predecessor. (Lu et al., 2006)

### *2.3. Stacking*

Stacking combines predictions from learner models with meta-models to create a final model with accurate predictions. The main advantage of a stacking ensemble is that it can protect the capabilities of various high-performing models used to solve various classification and regression problems. It also aids in preparing a better model with better predictions than all individual models. (Rajagopal et al., 2020)

## 3. Materials and Methods

### *3.1. Data Sets used for the Study*

For the analysis, two different borrower segments are considered: individual and enterprise borrowers. Individual borrowers take loans on their individual capacity, whereas enterprise borrowers are issued loans on their enterprise capacity. Both the data are collected by signing NDA. Individual borrower data is collected from a private bank in India, and enterprise data is collected from NBFC in India.

The individual data set has 105163 records, of which 100497 are negative class, and 4665 are positive class records. The pictorial distribution of the target distribution for the data set is given in Fig 2.

The enterprise data set has 97451 records, out of which 92900 are negative class and 4550 are positive class records. The pictorial representation of the target distribution for the data set is given in Fig 3.

Variables used for the analysis are shown in Table 1 and Table 2. Risk is a binary target variable, and others are independent variables. Both categorical and numerical variables are present in the dataset.

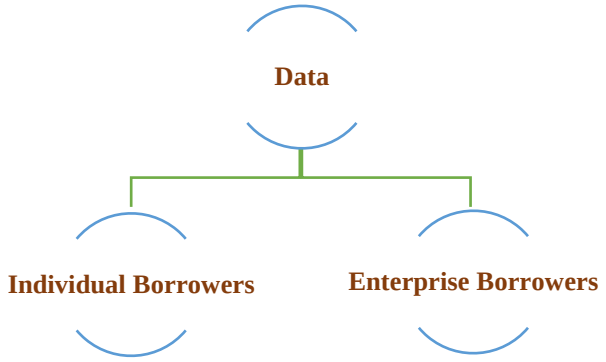


**Fig. 1 Borrower Segments**

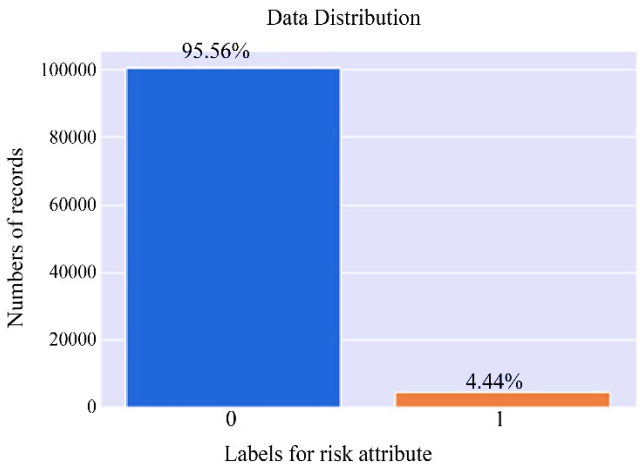


**Fig. 2 Target label percentage dataset 1**

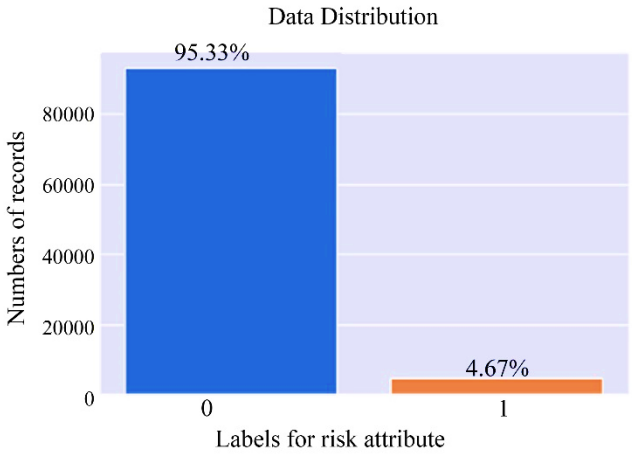


**Fig. 3 Target label percentage dataset 2**

**Table 1. Variable Description for Dataset 1**

| Variables | Description | Variable Type |
|---|---|---|
| Risk | The dependent binary variable represents whether an individual defaults or not. 0 and 1 denote the values for the non-defaulted and defaulted individuals, respectively. | Target |
| Gender | An independent variable represents the gender of the customer. | Categorical |
| Age | An independent variable represents the age of the customer. | Categorical |
| Marital status | An independent variable representing marital status. | Categorical |
| Qualification | An independent variable representing qualification. | Categorical |
| Employment | An independent variable representing employment. | Categorical |
| Loan Cycle No | An independent variable representing loan cycle number. | Numerical |
| Loan Period | An independent variable representing the loan period. | Numerical |
| Interest Rate | An independent variable representing interest rate. | Numerical |
| Next loan allowed | An independent variable represents whether the next loan is allowed or not. | Categorical |
| Number of Guarantor | An independent variable representing the number of guarantors. | Numerical |
| Guarantor 1 relation | An independent variable representing the relationship with the guarantor1. | Categorical |
| Ration Card Type | An independent variable representing the type of ration card | Categorical |
| Aadhar card | An independent variable representing whether an Aadhar card is present or not. | Categorical |
| Election card | An independent variable representing whether an election card is present or not. | Categorical |
| Pan card | An independent variable representing whether a pan card is present or not. | Categorical |
| Other account Bank Type | An independent variable representing accounts in other banks. | Categorical |
| HMF Policy Status | An independent variable representing HMF Policy Status | Categorical |
| Loan Product | An independent variable representing Loan Product | Categorical |
| Loan Reason | An independent variable representing Loan Reason | Categorical |
| NOMINEE_REL_TYP | An independent variable representing the nominee relationship. | Categorical |
| NOMINEE_AGE | An independent variable represents the nominee's age. | Numerical |
| MBR_RELIGION | An independent variable representing member religion. | Categorical |
| RD Account | An independent variable representing whether the RD account is present or not. | Categorical |
| Pension Account | An independent variable representing whether a pension account is present or not. | Categorical |
| loantoincomeratio | An independent variable representing the ratio of total loan to total income. | Numerical |
| percapitaincome | An independent variable represents percapta income. | Numerical |
| percapitaloan | An independent variable representing percapta loan | Numerical |
| percapitaexpenses | An independent variable represents percapta expenses. | Numerical |
| loantodepositratio | An independent variable represents the total loan ratio to the total deposit. | Numerical |
| amountduetoloanoutstanding | An independent variable representing the ratio of the total amount due to the total loan outstanding. | Numerical |
| expensetoincomeratio | An independent variable representing the ratio of total expenses to total income. | Numerical |
| amountduetofamilyincome | An independent variable represents the ratio of the total amount due to total family income. | Numerical |
| debt to income ratio | An independent variable representing the ratio of total debt to total income. | Numerical |

**Table 2. Variable Description for Dataset 2**

| Variables | Description | Variable Type |
|---|---|---|
| Risk | The dependent binary variable represents whether an MSME defaults or not. 0 and 1 denote the values for the non-defaulted and defaulted MSMEs, respectively. | Target |
| Sex | An independent variable representing the sex of a customer. | Categorical |
| Age | An independent variable representing the age of a customer. | Categorical |
| Loanamt | An independent variable representing loan amount. | Numerical |
| Tenure | An independent variable representing loan tenure. | Numerical |
| Yrresid | An independent variable representing the number of years at residence. | Numerical |
| Depndts | An independent variable representing the number of dependents. | Numerical |
| Bustype | An independent variable representing the type of business. | Categorical |
| Yrbusiness | An independent variable representing the number of years at the business. | Numerical |
| Busset | An independent variable representing business setup. | Categorical |
| NTC_cust | An independent variable representing new to credit customers or not. | Categorical |
| Location | An independent variable representing the location of MSME. | Categorical |
| Obligation | An independent variable representing obligations. | Numerical |
| Empoys | An independent variable representing employees. | Numerical |
| Busarea | An independent variable representing the business area. | Numerical |
| Income | An independent variable represents the income of MSME. | Numerical |
| Expenses | An Independent variable represents expenses. | Numerical |
| Highestloan | An independent variable representing the highest loan taken by the MSME. | Numerical |
| Liveloan | An independent variable representing liveloans. | Numerical |
| PACOut | An independent variable is represented per account outstanding. | Numerical |
| Pcloutloan | An independent variable represents the percapta outstanding loan. | Numerical |
| PCbankBC | An independent variable representing percapta bank branches. | Numerical |
| PCoutOffic | An independent variable representing percapta outstanding to the number of offices. | Numerical |
| debtratio | An independent variable representing debtratio. | Numerical |

### *3.2. Methodology*

Raw data is collected from 2 different sources, as mentioned in section 3.1. Various types of data preprocessing are done, and finally structured labelled dataset is generated. The complete research process is shown in Figure 4. To handle categorical variables present in the dataset, dummy encoding is done. All missing values are replaced with the median of the dataset, and correlation checking is done to ensure non-collinearity between independent variables. 80% of the total data is used for training and 20% for testing. The stacking approach is used for assembling with heterogeneous algorithms and multiple levels. Ten-fold cross-validation is carried out to avoid overfitting the algorithm. Various steps followed in the research process are

1. Data collection
2. Data preprocessing
3. Train -Test Split
4. Stacking Ensemble Classifier Implementation
5. Performance Analysis

## 4. Results and Discussion

The proposed algorithm has shown significant improvement in precision, recall, f1 score and accuracy on both datasets. Various machine learning algorithms are compared with the proposed model, namely Gaussian NB, Logistic Regression, Extra Trees Classifier, Random Forest Classifier, XGB Classifier, LGBM Classifier, and Neural Network. Machine learning algorithms' performance is enhanced by optimizing hyperparameters. These optimized best models are compared with the proposed algorithm. Comparative analysis of precision, recall, f1 score and accuracy are shown in Table 3 for dataset 1 and Table 4 for dataset 2. The experimental results presented in the table clearly show the proposed model's effectiveness against other machine learning algorithms.

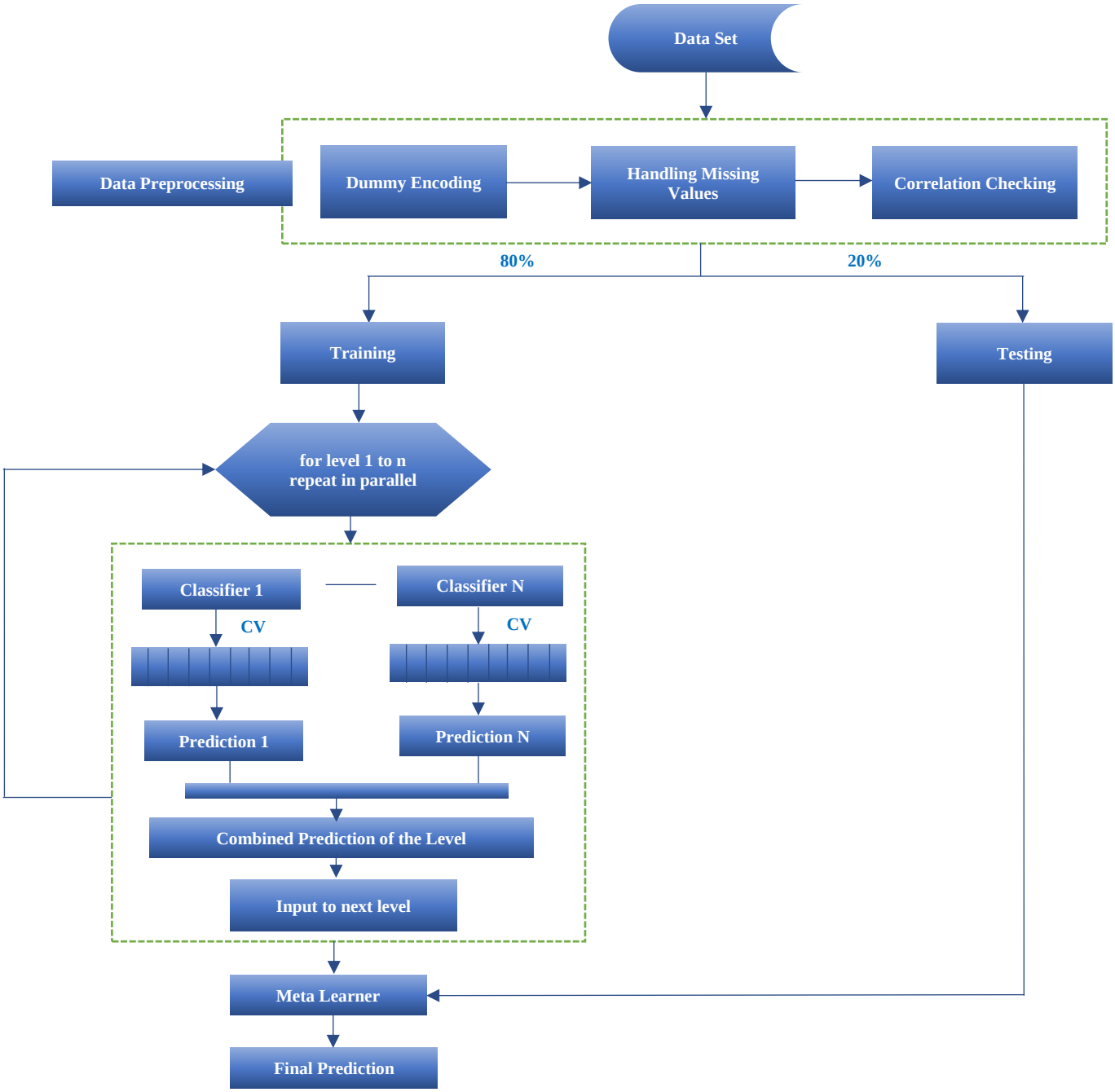


**Fig 4. Methodology for the proposed research**

**Table 3. Comparison of Proposed Ensemble Algorithm vs ML Algorithms (Dataset 1)**

| Algorithm | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Gaussian NB | 93.7% | 85.8% | 89.2% | 86.3% |
| Logistic Regression | 95.4% | 96.4% | 95.7% | 96.6% |
| Extra Trees Classifier | 94.5% | 87.7% | 89.8% | 96.2% |
| Random Forest Classifier | 96.2% | 96.3% | 95.4% | 96.9% |
| XGB Classifier | 97.3% | 97.2% | 96.5% | 97.1% |
| LGBM Classifier | 96.1% | 96.6% | 95.3% | 97.0% |
| Neural Network | 91.8% | 95.5% | 93.9% | 96.4% |
| Proposed Algorithm | **99.8%** | **99.8%** | **99.8%** | **99.8%** |

**Table 4. Comparison of Proposed Ensemble Algorithm vs ML Algorithms (Dataset 2)**

| Algorithm | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Gaussian NB | 93.7% | 91.8% | 92.2% | 91.3% |
| Logistic Regression | 93.4% | 95.4% | 93.7% | 95.6% |
| Extra Trees Classifier | 94.5% | 91.7% | 92.8% | 95.2% |
| Random Forest Classifier | 96.2% | 96.3% | 95.4% | 95.9% |
| XGB Classifier | 97.3% | 97.2% | 96.5% | 97.1% |
| LGBM Classifier | 96.1% | 96.6% | 95.3% | 98.0% |
| Neural Network | 96.8% | 96.5% | 97.9% | 95.4% |
| Proposed Algorithm | **99.9%** | **99.9%** | **99.9%** | **99.9%** |

## 6. Conclusion

The research paper introduced a novel ensemble algorithm for credit risk prediction. The proposed ensemble algorithm outperformed other ML algorithms, namely Gaussian NB, Logistic Regression, Extra Trees Classifier, Random Forest Classifier, XGB Classifier, LGBM Classifier and Neural Network. Performance metrics, namely Precision, Recall, F1-score and accuracy, are used for analyzing the performance of the proposed algorithm.

The precise interpretation of ensemble variety is still a mystery, despite the new method offering interesting insights into ensemble learning.

It still has to be disputed whether it can be explicitly defined or can only be understood intuitively. Ensemble diversity requires more investigation in a subsequent study. The sole focus on binary classification problems, the core component of the credit-risk evaluation, is another shortcoming of this research. Multi-class classification tasks should be applied using this new method, and appropriate studies should be designed in the future.

## Acknowledgement

The authors are grateful to a private bank and NBFC from India for providing data support to the study.

## References

[1] Wei-Yang Lin et al., “Machine Learning in Financial Crisis Prediction: A Survey,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 4, pp. 421–436, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[2] Shawni Dutta et al., “A Hybrid Machine Learning Model for Bank Customer Churn Prediction,” *International Journal of Engineering Trends and Technology*, vol. 70, no. 6, pp. 13–23, 2022. [CrossRef] [Publisher Link]

[3] Swati Warghade, Shubhada Desai, and Vijay Patil, "Credit Card Fraud Detection from Imbalanced Dataset Using Machine Learning Algorithm," *International Journal of Computer Trends and Technology*, vol. 68, no. 3, pp. 22-28, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Anton Markov, Zinaida Seleznyova, and Victor Lapshin, “Credit scoring methods: Latest Trends and Points to Consider,” *The Journal of Finance and Data Science*, vol. 8, pp. 180–201, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Arijit Bhattacharya, Saroj Kr. Biswas, and Ardhendu Mandal, “Credit Risk Evaluation: A Comprehensive Study,” *Multimedia Tools and Applications*, pp. 1–51, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Sudhansu R. Lenka, Bikram K. Ratha, and Biswaranjan Nayak, “A Review on Novel Approach to Handle Imbalanced Credit Card Transactions,” *International Journal of Engineering Trends and Technology*, vol. 62, no. 2, pp. 80-95, 2018. [CrossRef] [Publisher Link]

[7] Salvatore Carta et al., “Credit Scoring by Leveraging an Ensemble Stochastic Criterion in a Transformed Feature Space,” *Progress in Artificial Intelligence*, vol. 10, no. 4, pp. 417–432, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mayorga Lira Sergio Dennis, Laberiano Andrade-Arenas, and Miguel Angel Cano Lengua, "Credit Risk Analysis: Using Artificial Intelligence in a Web Application," *International Journal of Engineering Trends and Technology*, vol. 71, no. 1, pp. 305-316, 2023. [CrossRef] [Publisher Link]

[9] Wenyu Zhang, Dongqi Yang, and Shuai Zhang “A New Hybrid Ensemble Model with Voting-Based Outlier Detection and Balanced Sampling for Credit Scoring,” *Expert Systems with Applications*, vol. 174, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Wenyu Zhang et al., “A Novel Multi-Stage Ensemble Model with Enhanced Outlier Adaptation for Credit Scoring,” *Expert Systems With Applications*, vol. 165, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Yilun Jin et al., “A Novel Multi-Stage Ensemble Model with a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data,” *IEEE Access*, vol. 9, pp. 143593–143607, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jianrong Yao et al., “Novel Hybrid Ensemble Credit Scoring Model with Stacking-Based Noise Detection and Weight Assignment,” *Expert Systems with Applications*, vol. 198, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Himanshu Thakur, and Aman Kumar Sharma, "Supervised Machine Learning Classifiers: Computation of Best Result of Classification Accuracy," *International Journal of Computer Trends and Technology*, vol. 68, no. 10, pp. 1-8, 2020. [CrossRef] [Publisher Link]

[14] Abhijit Das, Pramod, "Exploratory Analysis on Anomaly-based IDS Data Using DASK and Ensemble Learning: A Data Parallelization Approach," *International Journal of Engineering Trends and Technology*, vol. 70, no. 12, pp. 370-391, 2022. [CrossRef] [Publisher Link]

[15] Indu Singh et al., "A Multi-level Classification and Modified PSO Clustering Based Ensemble Approach for Credit Scoring," *Applied Soft Computing*, vol. 111, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] S. Yamashkin et al., "Classification of Metageosystems by Ensembles of Machine Learning Models," *International Journal of Engineering Trends and Technology*, vol. 70, no. 9, pp. 258–268, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] A. Lakshmanarao, and M. Shashi, "An Efficient Android Malware Detection Framework with Stacking Ensemble Model," *International Journal of Engineering Trends and Technology*, vol. 70, no. 4, pp. 294–302, 2022. [CrossRef] [Publisher Link]

[18] Zarapala Sunitha Bai, and Sreelatha Malempati, "An Enhanced Text Mining Approach using Ensemble Algorithm for Detecting Cyber Bullying," *International Journal of Engineering Trends and Technology*, vol. 70, no. 9, pp. 393–399, 2022. [CrossRef] [Publisher Link]

[19] Diwakar Tripathi et al., "Credit Scoring Models Using Ensemble Learning and Classification Approaches: A Comprehensive Survey," *Wireless Personal Communications*, vol. 123, no. 1, pp. 785–812, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] G.Gowrishankar, S.Balgani, and R.Aruna, "Multiselfish Attacks Detection Based on Credit Risk Information in Cognitive Radio Ad-hoc Networks," *SSRG International Journal of Computer Science and Engineering*, vol. 2, no. 4, pp. 1-7, 2015. [CrossRef] [Publisher Link]

[21] Sudhansu R. Lenka et al., "Empirical Analysis of Ensemble Learning for Imbalanced Credit Scoring Datasets: A Systematic Review," *Wireless Communications & Mobile Computing*, vol. 2022, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Akinbohun Folake, Akinbohun Ambrose, and E. Oyinloye Oghenerukevwe, "Stacked Ensemble Model for Hepatitis in Healthcare System," *International Journal of Computer and Organization Trends*, vol. 9, no. 4, pp. 25-29, 2019. [CrossRef] [Publisher Link]

[23] Q. Sun, and B. Pfahringer, "Bagging Ensemble Selection," *Advances in Artificial Intelligence*, vol. 7106, pp. 251–260, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[24] J. Lu et al., "Ensemble-Based Discriminant Learning with Boosting for Face Recognition," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 166–178, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[25] Smitha Rajagopal, Poornima Panduranga Kundapur, and Katiganere Siddaramappa Hareesha, "A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets," *Security and Communication Networks*, vol. 2020, 2020. [CrossRef] [Google Scholar] [Publisher Link]