International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

# A Novel Ensemble-Deep Learning Approach for Accurate Credit Risk Prediction in Imbalanced Financial Datasets

**Rathan Kumar Chenoori[1], Sunil Kumar Thota[2], Pillareddy Vamsheedhar Reddy[3], Padma BalaKrishna[4]**

1,2 Assistant Professor, Department of CSE, Keshav Memorial Institute of Technology, Hyderabad, Telangana, India

3 Assistant Professor, Department of CSE-AIML, Keshav Memorial Engineering College, Hyderabad, Telangana, India

4 Assistant Professor, Department of CSE, Keshav Memorial Engineering College, Hyderabad, Telangana, India

**Abstract:** The accurate prediction of credit risk is a crucial endeavor in the banking and finance sector which is impeded by the complexity and inaccessibility of financial data. This paper will use the Australian and German Credit data, and the preprocessing methods that will be applied are the management of class imbalance through SMOTEENN and feature extraction to identify relevant features. Various machine learning and deep learning models such as Convolutional Neural Networks (CNN), Multi-Layer Perceptron (MLP), Random Forest, and Logistic Regression are analyzed and evaluated using different performance indicators such as accuracy, precision, recall, F1-score, sensitivity, specificity, and confusion matrices. The approach suggested is a Stacking Classifier, which has strong generalization and predictive power in both data sets. A Voting Classifier (which combines Bagging with RF and AdaBoost with DT) is used to increase the accuracy, therefore, increasing the resilience and the generalization as a whole. The Voting Classifier achieves 100 percent accuracy on the German dataset (original and resampled), 98.3 percent on the resampled Australian dataset and 89.1 percent on original Australian dataset. To understand the model predictions and determine the significance of features, explainable AI approaches like LIME and SHAP are applied, therefore, improving the transparency and reliability. The models and preprocessing artifacts are finally made available through a Flask web application, where users can make real-time credit risk predictions with help of interpretability.

***"Index Terms:*** *Credit risk, CNN, ensemble learning, machine learning, MLP, random forest".*

## 1. INTRODUCTION

Financial risk management is based on credit risk prediction, which has a significant influence on the profitability and stability of lending institutions all over the world. Poor credit scoring does not only expose banks to losses but also poor credit scoring weakens market trust and resilience. Determining the trusted borrowers and potential defaulters is the key to successful financial operations.

Traditional statistical procedures, though basic, often fail to reflect the complexity of modern financial systems. The credit risk is influenced by numerous factors such as the macroeconomic changes, the behavior of borrowers and market volatility and hence the rule-based models are not sufficient to address such complexities [4]. This has led to a growing need by financial institutions to have data-driven and dynamic structures that can support such complex interconnections [5].

ML has emerged as a powerful tool to overcome these limitations, offering the ability to process large, heterogeneous data sets and detect subtle trends linked to creditworthiness [6]. Unlike the traditional approach, ML models can be used to examine past patterns and apply the knowledge to new applicants more quickly and accurately [7]. The approaches based on logistic regression to advanced neural networks have demonstrated considerable credit scoring system improvements, placing ML at the centre of the impactful change to financial analytics [8].

Despite these developments, there remains a significant challenge: the unfair nature of credit data. In many loan cases, the ratio of defaulters to the applicants is only a small percentage, which leads to biased predictions that are favorable to the majority group [9]. This gap prevents the detection of high-risk borrowers, which is critical in reducing the financial losses. Scholars have explored data resampling techniques that create equal training samples, which increases sensitivity of the model to the minority group [10].

The given paper presents a more advanced ensemble architecture, which includes the implementation of multiple ML models and more sophisticated resampling methods. The goal of this method is to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

combine different classifiers in order to find linear and non-linear relationships in financial data as well as address the issue of imbalance in datasets. The study discusses three main questions; how stacked ensembles improve predictive reliability compared to single models, the effectiveness of hybrid resampling methods in detecting minority classes, and the relative classification performance of the method used to the existing models.

## 2. LITERATURE REVIEW

The challenge of estimating credit risk when the data may be unevenly distributed has acquired much interest during the past few years, and several scholars have suggested new approaches to increase the quality of classification and generalization. Jiang, Lu, Wang, and Ding [11] conducted an important benchmarking research in order to evaluate the state-of-the-art unbalanced learning techniques in the context of credit scoring. They systematically tested oversampling, undersampling, and hybrid algorithms, as well as algorithmic approaches such as cost-sensitive learning, on different financial data. The authors concluded that some resampling methods, including SMOTE and its variations, can improve the sensitivity of minority classes, but no single strategy was found to outperform all other evaluation measures. This benchmarking highlighted the importance of dataset properties and model choice in credit scoring, and also stimulated the study of hybrid ensemble models with the ability to integrate multiple strategies in a dynamic way to provide better stability in performance.

A special study by Zhao, Cui, Ding, Li, and Bellotti [12] on resampling methodologies was aimed at solving the problem of class imbalance in credit risk prediction. They compared the effects of common resampling techniques, such as random oversampling, random undersampling, and hybrid approaches, such as SMOTE-ENN. The findings showed that hybrid approaches, which create instances of minority classes and at the same time clean up noisy majority samples, significantly enhance recall of default cases. However, they have admitted that the potential risk of overfitting could exist in the case of non-meticulous calibration of oversampling. Zhao et al. emphasized the importance of verifying the production of synthetic data and integrating resampling using the powerful classifiers to increase the generalization in practical use.

Alam, Shaukat, Hameed, Luo, Sarwar, Shabbir, Li, and Khushi [13] studied credit card default prediction using considerably unbalanced datasets, and offered empirical information on the effectiveness of ML classifiers at different degrees of imbalance. They examined decision trees, random forests, boosting algorithms, and neural networks using resampling methods and non-resampling. Ensemble based classifiers, including RF and AdaBoost, had shown competitive performance; however, they required resampling strategies in order to be able to accurately determine the patterns of minority classes. The authors highlighted that the degree of imbalance significantly affects the performance of the model, which is why the combination of preprocessing procedures with advanced classification models is necessary to guarantee the successful outcome of the model.

Wang [14] suggested an unbalanced credit risk forecasting model which is based on SMOTE multi-kernel fuzzy c-means (FCM) clustering algorithm optimized by particle swarm optimization (PSO). This mixed approach not only corrected the imbalance in the classes but also improved the quality of clustering, which enabled better data representation prior to classification. The methodology provided by Wang disclosed that the combination of optimization procedures with resampling schemes could enlarge the feature groupings as well as the classifier effectiveness, thus, leading to the recognition of defaulters in unbalanced data. The work contributed to the existing literature by showing how optimization-based clustering can be used to supplement traditional resampling to provide better predictability of risks.

Despite the differences in their areas of focus, studies on DL structures have produced cross-cutting solutions to feature extraction and hybrid models. Amin, Alsulaiman, Muhammed, Mekhtiche, and Hossain [15] proposed a multi-layer CNN feature fusion approach to EEG motor imagery classification that demonstrates the effectiveness of hierarchical convolutional feature extraction to obtain complex, high-dimensional features. Their approach proved the importance of convolutional layers in feature fusion, which is applicable to credit risk tasks, where financial data contains complicated interactions. Kour and Gupta [16] created a hybrid DL model to predict depression using Twitter data by combining CNNs to extract features with

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

bidirectional LSTMs to learn a sequence. This hybrid architecture proved that a combination of complementary neural architectures can be effective in capturing local and sequential dependencies, which can also be applied to temporal financial data in credit risk analysis. All these studies highlight the effectiveness of CNNs and hybrid neural networks as base learners in stacked ensembles in financial forecasting tasks.

Alam et al. [17] carried out an in-depth investigation of imbalance problem by providing an empirical examination of credit card default forecasting with the use of a variety of machine learning algorithms on unbalanced data. Their study indicated that resampling and feature selection are important in improving prediction accuracy when the imbalance ratios vary. Alam and the others found that without preprocessing, the models failed to generalize to the minority classes, but when hybrid preprocessing procedures were used, recall increased significantly. Their results supported past claims made by Zhao et al. [12] about the importance of resampling design, but also added to the discussion on how feature engineering, resampling and classifier selection interact to affect final results.

Besides the predicted accuracy, interpretability has emerged as an important feature of credit risk assessment. Suhadolnik, Ueyama and Da Silva [18] explored machine learning architecture in order to enhance credit risk assessment with a focus on interpretability and practicality. Their practical study emphasized the importance of transparent models that can provide information about decision-making and thereby satisfy the regulatory requirements in the financial sector. They explored interpretable model families with post-hoc explanations, demonstrating how the use of techniques such as feature importance analysis may improve trust and acceptance of AI systems in finance. This work corresponds with the growing use of explainable AI techniques such as LIME and SHAP into credit scoring systems.

Classical algorithm study is also an important research in this field. Shilpa, Shaha, Hajek, and Abedin [19] constructed default risk prediction models based on the SVM and the logit-support version of the SVM. Their results suggested that margin-based classifiers are as competitive as ensemble models particularly when carefully tuned using kernel functions and class-weighting processes to eliminate imbalance. The authors demonstrated that SVM-based methods can serve as useful baselines and modules within ensemble stacks, which can be strong in terms of performance and interpretability through combination with other algorithms.

Mienye and Jere [20] carried out an extensive review of decision trees, their principles, methods, and their applications. Their evaluation emphasized how decision trees could be flexibly used to handle various feature space, how they integrated well with ensemble methods, such as bagging and boosting, and how they continued to be important in fields such as credit scoring. Mienye and Jere have noted the strengths and weaknesses of decision tree methods, especially regarding their role in financial analytics as standalone classifiers and as the starting point in ensemble systems.

## 3. MATERIALS AND METHODS

The suggested solution focuses on accurate forecasting of credit risk using the Australian and German Credit data based on the implementation of advanced ML and DL techniques. The preprocessing of the data includes handling the missing data, coding the categorical variables through the use of LabelEncoder [21], and normalization of the numerical features using StandardScaler [22]. To correct the imbalance between the classes, SMOTE-ENN is used to ensure that the classes are fairly represented [26]. The feature extraction reduces the dimensions and enhances the model efficiency [27]. CNN [29], MLP [24], RF [23], and LR [28] are representative predictive models that are capable of capturing diverse patterns of data. Ensemble techniques are performed through a Stacking Classifier, which combines CNN, Random Forest, and Logistic Regression and through a Voting Classifier, which combines Bagging with RF and AdaBoost with Decision Tree. Explainable AI procedures LIME and SHAP enhance interpretability, and the computation is simplified with the help of a Flask-based web application to make real-time predictions [30].

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

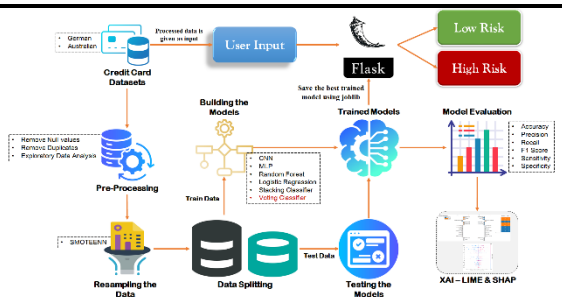



Fig.1 Proposed Architecture

The suggested system design (fig.1) outlines a complete credit risk forecasting pipeline based on German and Australian credit data. The initial stage of data preprocessing involves the null values removal, deletion of duplicates, and the implementation of the exploratory data analysis. SMOTEENN deals with class imbalance, which is followed by data partitioning into training and testing set. Different ML and DL systems, like CNN, MLP, RF, LR and ensemble classifiers, are trained and evaluated with the help of large performance measures. The best model is deployed through Flask interface, which enables real-time risk prediction with explanation through LIME and SHAP [25].

**a) Dataset Collection:**

***German Dataset:*** The dataset (German credit) comprises 1,000 items with 10 attributes, such as Age, Sex, Job, Housing, Savings accounts, Checking account, Credit amount, Duration and Purpose. Categorical variables include Sex, Housing, Savings accounts, Checking accounts, and Purpose which have been coded to prepare the model. The value of the credit is used to create a binary tag based on the median. The data has been cleared of any null or duplicate records, and indexes are set to zero. This is a fined and balanced dataset which enables accurate prediction of credit risk after resampling with SMOTEENN.

| | Unnamed: 0 | Age | Sex | Job | Housing | Saving accounts | Checking account | Credit amount | Duration | Purpose |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 67 | male | 2 | own | unknown | little | 1169 | 6 | radio/TV |
| 1 | 1 | 22 | female | 2 | own | little | moderate | 5951 | 48 | radio/TV |
| 2 | 2 | 49 | male | 1 | own | little | unknown | 2096 | 12 | education |
| 3 | 3 | 45 | male | 2 | free | little | little | 7882 | 42 | furniture/equipment |
| 4 | 4 | 53 | male | 2 | free | little | little | 4870 | 24 | car |

Fig.2 German Dataset

***Australian Dataset:*** The Australian credit data set consists of 690 items and 14 features (named A1 to A14) and the target variable. Features include quantitative and qualitative financial aspects including account balances, work status, age, and credit term. There are no cases of missing or duplication of data and categorical columns have been handled to conform to the model. The target variable is the credit risk. After preprocessing and SMOTEENN resampling, the data is balanced, which will form a reliable foundation to train and test models and distinguish between high-risk and low-risk applicants.

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | Target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22.08 | 11.46 | 2 | 4 | 4 | 1.585 | 0 | 0 | 0 | 1 | 2 | 100 | 1213 | 0 |
| 1 | 0 | 22.67 | 7.00 | 2 | 8 | 4 | 0.165 | 0 | 0 | 0 | 0 | 2 | 160 | 1 | 0 |
| 2 | 0 | 29.58 | 1.75 | 1 | 4 | 4 | 1.250 | 0 | 0 | 0 | 1 | 2 | 280 | 1 | 0 |
| 3 | 0 | 21.67 | 11.50 | 1 | 5 | 3 | 0.000 | 1 | 1 | 11 | 1 | 2 | 0 | 1 | 1 |
| 4 | 1 | 20.17 | 8.17 | 2 | 6 | 4 | 1.960 | 1 | 1 | 14 | 0 | 2 | 60 | 159 | 1 |

Fig.3 Australian Dataset

**b) Exploratory Data Analysis (EDA):**

The purpose of the EDA is to understand the organization, distribution, and patterns of data sets. This involves the analysis of data types, identification of categorical and numerical attributes, checking the existence of the null value and checking the distributions of the classes. The imbalances in target classes are explained by visualisation tools, such as count charts. Exploratory Data Analysis can be used to detect anomalies, patterns and interdependence between characteristics and therefore informs feature selection and preprocessing. It supports informed decision-making at later stages, which improves the model performance and explainability to assess credit risk.

**c) Pre-processing:**

Pre-processing involves cleaning the raw data and transforming it into a model acceptable by ML. Some of the steps include dealing with missing values, removing duplicates, encoding the category variables using techniques like Label Encoding and normalizing or scaling the numerical variables. Additional tasks include the creation of new labels, feature extraction and ensuring consistency between datasets. These transformations reduce noise, solve discrepancies, and make the data machine-readable, making them more predictive and stable and allowing them to interact with resampling and model training strategies.

**d) Data Imbalancing and Balancing:**

Financial records often indicate imbalance in classes, which is in that there is a low number of high-risk applicants in comparison with the low-risk applicants. To overcome this, the technique of resampling, like SMOTEENN, is used, which combines the oversampling of minority classes with the removal of the noisy examples. This balances the data set, which enhances the learning performance of machine learning models. Balanced datasets

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

eliminate the bias of majority classes, increase the generalization of models and ensure that there is fair evaluation of accuracy and precision, recall, and other measures, resulting in reliable and robust credit risk predictions in diverse applicant groups.

**e) Training and Testing:**

After preprocessing and balancing, the two datasets are split into training and testing data to ensure that an unbiased evaluation is achieved. The models identify trends based on the sets of features in order to differentiate high-risk and low-risk applicants. Fair datasets facilitate learning in that they reduce biasness on dominant classes. Testing models with new data will ensure that they can be generalized, and therefore give reliable predictions. Resampled datasets help the models to detect subtle differences between the applicants as well as provide consistency, stability and strength to the credit risk prediction process.

**f) Algorithms:**

**Convolutional Neural Network (CNN):** Learns hierarchical feature representations of the input data, which capture complex patterns and associations. The convolution and pooling layers unearth some meaningful relationships among credit data, which improves the classification rates and detects subtle differences between risky and non-risky scenarios.

$$S(i,j) = \sum_{m}\sum_{n} I(i+m, j+n)\cdot K(m,n) \quad (1)$$

**Multi-Layer Perceptron (MLP):** A feedforward neural network, which denotes non-linear correlations between features and results. Latent features detect meaningful patterns and activation functions introduce non-linearity, which enables them to accurately differentiate between risky and non-risky credit cases and generalize to multiple financial data sets.

$$\hat{y} = f\left(W^{L} f\left(W^{L-1} \cdots f\left(W^{1}X + b^{1}\right) + b^{(L-1)}\right) + b^{L}\right) \quad (2)$$

**Random Forest (RF):** A set of decision trees which combines numerous models to enhance precision in categorization. It reduces the overfitting, handles high-dimensional data, and evaluates subsets of features to accurately distinguish between high-risk and low-risk credit applicants and provides an insight into significant aspects.

$$Gini = 1 - \sum_{i=1}^{C} (Pi)^2 \quad (3)$$

**Logistic Regression (LR):** Predicts probabilities of binary variables with the use of financial and demographic variables. Offers decipherable coefficients to explain the features impact, provides a reliable benchmark with which to compare, and enables a clear and convincing assessment of high and low-risk credit applicants.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(W^{T}x + b)}} \quad (4)$$

**Stacking Classifier:** Combines predictions of the different underlying models, such as CNN, LR, and RF, to determine different patterns. Enhances the overall predictive effectiveness, fills deficiencies of the individual models, balances the trade-offs between bias and variance, and produces robust and accurate credit risk predictions.

$$\hat{y} = g(Y_{base}) = g\left(f_1(x), f_2(x), \ldots, f_m(x)\right) \quad (5)$$

**Voting Classifier:** Majority vote or weighted averages Predictions, e.g. Bagging with RF and AdaBoost with Decision Tree, are combined by majority vote or average weights. Guarantees increase accuracy and robustness through datasets and minimize the effect of the single model errors.

$$\hat{y} = argmax_{c}\left(\sum_{i=1}^{n} II\ (\hat{y}_i = c)\right) \quad (6)$$

**g) Integration of XAI and Flask Framework**

The proposed solution will be based on the integration of XAI tools and the Flask web app to provide interpretable and real-time credit risk judgments. LIME comes up with localized explanations to the particular predictions and it focuses on the influence of each attribute to the projected outcome. SHAP improves this by showing global and local interpretability through summary plots, dependence plots, decision plots, and waterfall visualizations to demonstrate the effect of parameters like Age, Credit amount, and account balances on risk scores. Such explanations increase the level of transparency and allow the users to understand the logic of the model and develop trust in the automated credit decisions.

Flask is a bare-bones web application on which users can input demographic and financial information. Immediate predictions are provided with LIME and SHAP visuals, which are useful to conduct interactive and interpretable credit risk analysis. This integration ensures that it is practically usable and provides easy understanding of the importance of the features and decision making hence making the system suitable to both the technical and non-technical users.

## 4. EXPERIMENTAL RESULTS

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

**Accuracy:** The test accuracy is related to the ability of a test to differentiate patient and health cases correctly. To determine the validity of a test, there is a need to calculate a ratio of true positives and true negatives on all the cases evaluated. This may be written in mathematical notation as:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} (7)$$

**Precision:** Precision measures the rate of correctly classified cases of those detected as positive. Therefore, the precision equation will be:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} (8)$$

**Recall:** Recall is a ML metric that determines the ability of a model to identify all relevant examples of a certain type. The proportion of the correct positive observations of the overall actual positives, which provide information about the effectiveness of a model in detecting the incidences of a given class.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} (9)$$

**F1-Score:** F1 score is a statistic that is employed to measure the precision of a ML model. It combines the recall and accuracy of a model. The accuracy measure is a measure of the correct prediction rate given by a model on the full set of data.

$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 (10)$$

**Sensitivity:** The sensitivity determines how effective a test or equipment is in revealing a condition in an individual. It is established through comparison of the number of people who test positive of an ailment and the real prevalence of the illness.

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (11)$$

**Specificity:** It is calculated by calculating the number of people who tested negative of something and then dividing it by the total of people who did not have the ailment, including people who tested negative and people who tested positive, but were not actually having the ailment.

$$Specificity = \frac{TN}{(TN + FP)} \quad (12)$$

***Table.1*** Performance Evaluation Table - Australian Dataset – Data Resampling

| ML Model | Accuracy | F1 Score | Recall | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| CNN | 0.917 | 0.917 | 0.917 | 0.917 | 0.917 | 0.918 |
| MLP | 0.983 | 0.983 | 0.983 | 0.984 | 0.983 | 0.982 |
| Random Forest | 0.967 | 0.967 | 0.967 | 0.969 | 0.967 | 0.969 |
| Logistic Regression | 0.983 | 0.983 | 0.983 | 0.984 | 0.983 | 0.984 |
| Stacking Classifier | 0.967 | 0.967 | 0.967 | 0.967 | 0.967 | 0.967 |
| **Voting Classifier** | **0.983** | **0.983** | **0.983** | **0.984** | **0.983** | **0.984** |

Table 1 shows that the Voting Classifier outperforms other models showing the greatest general predictive effectiveness and robust credit risk classification.

***Table.2*** Performance Evaluation Table – Australian Dataset – Original Data

| ML Model | Accuracy | F1 Score | Recall | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| CNN | 0.855 | 0.853 | 0.855 | 0.854 | 0.855 | 0.810 |
| MLP | 0.710 | 0.705 | 0.710 | 0.716 | 0.710 | 0.725 |
| Random Forest | 0.877 | 0.878 | 0.877 | 0.882 | 0.877 | 0.839 |
| Logistic Regression | 0.862 | 0.862 | 0.862 | 0.862 | 0.862 | 0.846 |

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

| | | | | | | |
|---|---|---|---|---|---|---|
| Stacking Classifier | 0.855 | 0.855 | 0.855 | 0.854 | 0.855 | 0.842 |
| **Voting Classifier** | **0.891** | **0.892** | **0.891** | **0.894** | **0.891** | **0.863** |

The Voting Classifier has better performance in Table 2 whereby it outperforms other models in accurately predicting credit risk.

***Table.3*** Performance Evaluation Table – German Dataset – Data Resampling

| ML Model | Accuracy | F1 Score | Recall | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| CNN | 0.905 | 0.904 | 0.905 | 0.914 | 0.905 | 0.904 |
| MLP | 0.945 | 0.945 | 0.945 | 0.945 | 0.945 | 0.945 |
| Random Forest | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Logistic Regression | 0.940 | 0.940 | 0.940 | 0.941 | 0.940 | 0.939 |
| Stacking Classifier | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **Voting Classifier** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |

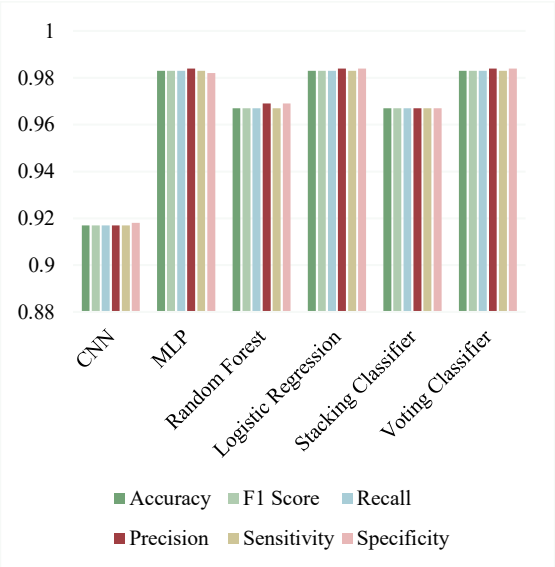
Table 3 shows that over the rest of the models, the Random Forest, Stacking Classifier, and Voting Classifier are much better in their performance of flawless credit risk prediction.

***Table.4*** Performance Evaluation Table – German Dataset – Original Data

| ML Model | Accuracy | F1 Score | Recall | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| CNN | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.961 |
| MLP | 0.925 | 0.925 | 0.925 | 0.926 | 0.925 | 0.921 |
| Random Forest | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Logistic Regression | 0.955 | 0.955 | 0.955 | 0.955 | 0.955 | 0.954 |
| Stacking Classifier | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **Voting Classifier** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |

Table 4 shows that the RF, Stacking Classifier, and Voting Classifier models outperform the rest and achieve a perfect prediction in credit risk assessment.

***Fig. 4*** Comparison Graph - Australian Dataset – Data Resampling




A comparative graph in figure 4 shows the effectiveness of some of the models, with the Voting Classifier as the better model in general.

***Fig. 5*** Comparison Graph – Australian Dataset – Original Data

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

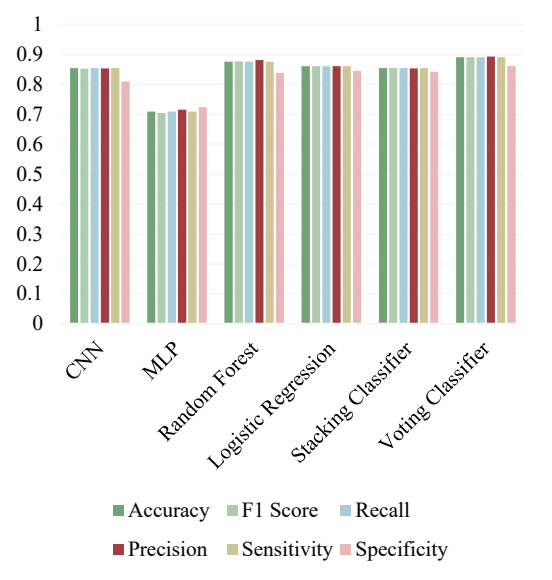
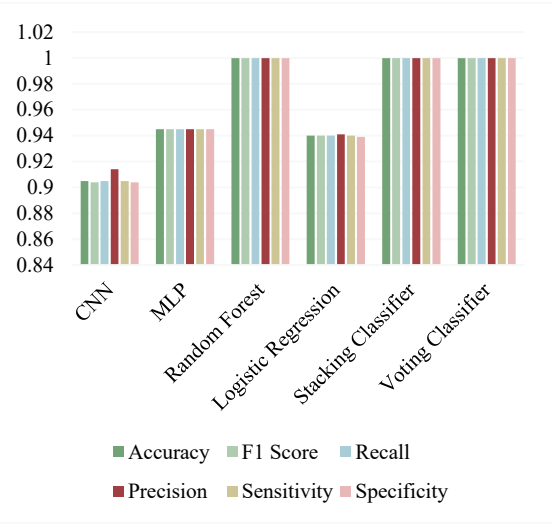



Fig. 5 shows that the Voting Classifier has the best performance, as it outperforms other models in terms of graphically depicting the effectiveness of credit risk prediction.

***Fig. 6*** Comparison Graph – German Dataset – Data Resampling




The Random Forest, Stacking Classifier, and Voting Classifier also outperform the rest of the models in Fig. 6, which is an absolute sign of the best effectiveness in credit risk predicting.

***Fig. 7*** Comparison Graph – German Dataset – Original Data

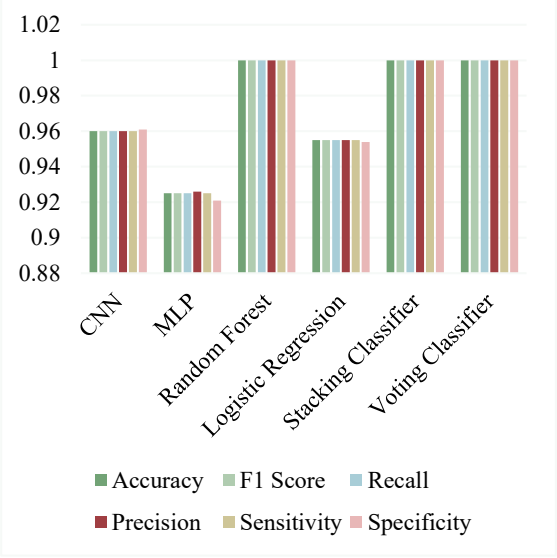



Fig. 7 shows that the RF, Stacking Classifier, and Voting Classifier outperform other models, and they have perfect predictive performance in credit risk analysis.

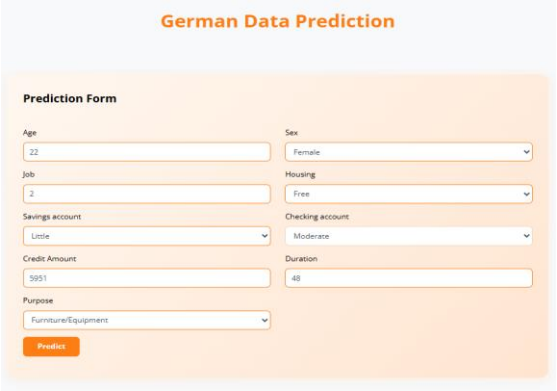



Fig. 8 Upload Input Data

Fig. 8 shows the input interface where the user is allowed to enter credit data such as account status, amount, duration and loan purpose before prediction.

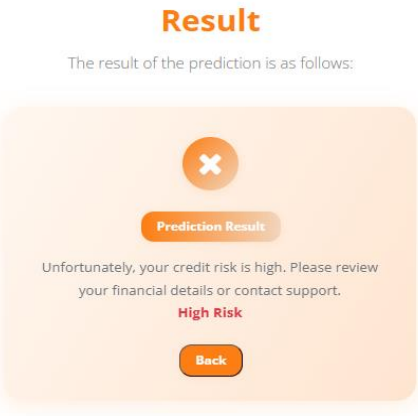



Fig. 9 Predicted Results

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

Figure 9 represents the type of input form that requests users to enter their account details, such as type of account, credit amount, term, and purpose of the loan to be used in predictive analysis.

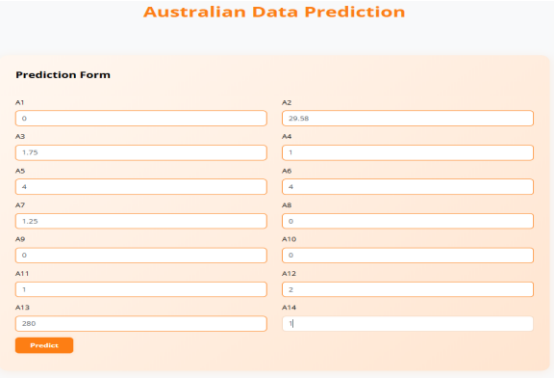



Fig. 10 Upload Input Data

In Figure 10, the screenshot of a user interface is provided to input the financial and demographic variables that are relevant to the prediction of credit risks.

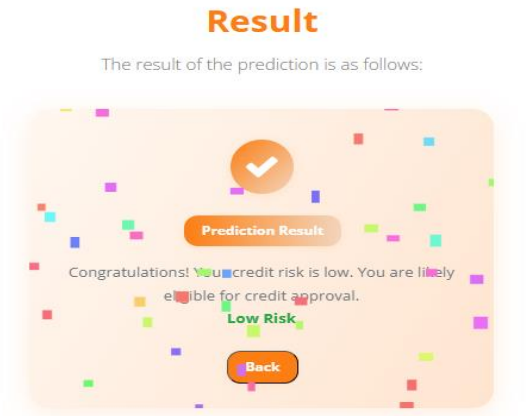



Fig. 11 Predicted Results

The output screenshot, which shows the expected credit risk, the respective probabilities, and also the major feature contributions to the applicant, appears in figure 11.

## 5. CONCLUSION

The developed credit risk prediction framework demonstrates the effectiveness of applying a complex ML and DL framework with data balancing and ensemble strategies to deal with problems in financial data. The Australian and German Credit dataset were preprocessed, this involved dealing with the issue of class imbalance with SMOTEENN, feature extraction, category encoding, and normalization of the data to facilitate modeling. CNN, MLP, RF and LR were trained and evaluated on the basis of accuracy, precision, recall, F1-score, sensitivity, specificity and confusion matrices. The proposed Stacking Classifier which combines CNN, LR and RF showed good generalization and good predictive ability on both datasets. A Bagging ensemble Voting Classifier with RF and AdaBoost with DT reached 100 percent on the German dataset (both original and resampled), 98.3 percent on the resampled Australian dataset, and 89.1 percent on the original Australian dataset. LIME and SHAP are explainable AI methodologies that clarified the importance of the features, thereby improving transparency and interpretability. The introduction of Flask web application allows real-time prediction of credit risks given the input of a user, demonstrating the reliability, stability, and usefulness of the system.

Credit risk prediction system can be improved by incorporating additional financial data to increase the model generalization and flexibility in a variety of banking situations. Predictability, forecast accuracy, can be enhanced by the use of real-time data streams and automatic data upgrades. Future research can be related to the implementation of more advanced DL models, such as attention-based ones, to explain complex associations between features. The deployment using cloud platforms improves scalability and accessibility to many users. By incorporating explainable AI to perform automated feature analysis and establish risk mitigation strategies, one can get actionable information, which increases the strength, transparency, and usefulness of the system in financial decision-making.

## REFERENCES

[1] Zhao, Z., Cui, T., Ding, S., Li, J., & Bellotti, A. G. (2024). Resampling techniques study on class imbalance problem in credit risk prediction. Mathematics, 12(5), 701.

[2] Kou, G., Chen, H., & Hefni, M. A. (2022). Improved hybrid resampling and ensemble model for imbalance learning and credit evaluation. Journal of Management Science and Engineering, 7(4), 511-529.

[3] Aruleba, I., & Sun, Y. (2025). Enhanced credit risk prediction using deep learning and SMOTE-ENN resampling. Machine Learning with Applications, 100692.

[4] Aruleba, I., & Sun, Y. (2024). Effective credit risk prediction using ensemble classifiers with model explanation. IEEE Access.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

[5] Mienye, I. D., & Sun, Y. (2023). A deep learning ensemble with data resampling for credit card fraud detection. Ieee Access, 11, 30628-30638.
[6] C.-F. Wu, S.-C. Huang, C.-C. Chiou, and Y.-M. Wang, ‘‘A predictive intelligence system of credit scoring based on deep multiple kernel learning,’’ Appl. Soft Comput., vol. 111, Nov. 2021, Art. no. 107668.
[7] Y. Shi, Y. Qu, Z. Chen, Y. Mi, and Y. Wang, ‘‘Improved credit risk prediction based on an integrated graph representation learning approach with graph transformation,’’ Eur. J. Oper. Res., vol. 315, no. 2, pp. 786–801, Jun. 2024.
[8] G, Viswanath., N, Madhvik., K, Bhaskar., K, Supriya. (2024). Machine-Learning-Based Cloud Intrusion Detection. International Journal of Mechanical Engineering Research and Technology, 16(9), 38-52.
[9] J. P. Noriega, L. A. Rivera, and J. A. Herrera, ‘‘Machine learning for credit risk prediction: A systematic literature review,’’ Data, vol. 8, no. 11, p. 169, Nov. 2023.
[10] S. Bhatore, L. Mohan, and Y. R. Reddy, ‘‘Machine learning techniques for credit risk evaluation: A systematic literature review,’’ J. Banking Financial Technol., vol. 4, no. 1, pp. 111–138, Apr. 2020.
[11] C. Jiang, W. Lu, Z. Wang, and Y. Ding, ‘‘Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring,’’ Expert Syst. Appl., vol. 213, Mar. 2023, Art. no. 118878.
[12] Z. Zhao, T. Cui, S. Ding, J. Li, and A. G. Bellotti, ‘‘Resampling techniques study on class imbalance problem in credit risk prediction,’’ Mathematics, vol. 12, no. 5, p. 701, Feb. 2024.
[13] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, ‘‘An investigation of credit card default prediction in the imbalanced datasets,’’ IEEE Access, vol. 8, pp. 201173–201198, 2020.
[14] Viswanath, G., Prasad, K. K., Lakshmi, J. M., & Swapna, G. (2025). Diabetes Diagnosis Using Machine Learning with Cloud Security. Cuestiones De Fisioterapia, 54(2), 417-431. https://doi.org/10.48047/r2mhn978
[15] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, ‘‘Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion,’’ Future Gener. Comput. Syst., vol. 101, pp. 542–554, Dec. 2019.
[16] H. Kour and M. K. Gupta, ‘‘A hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM,’’ Multimedia Tools Appl., vol. 81, no. 17, pp. 23649–23685, Jul. 2022.
[17] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, ‘‘An investigation of credit card default prediction in the imbalanced datasets,’’ IEEE Access, vol. 8, pp. 201173–201198, 2020.
[18] N. Suhadolnik, J. Ueyama, and S. Da Silva, ‘‘Machine learning for enhanced credit risk assessment: An empirical approach,’’ J. Risk Financial Manage., vol. 16, no. 12, p. 496, Nov. 2023.
[19] N. A. Shilpa, P. Shaha, P. Hajek, and M. Z. Abedin, ‘‘Default risk prediction based on support vector machine and logit support vector machine,’’ in Novel Financial Applications of Machine Learning and Deep Learning: Algorithms, Product Modeling, and Applications. Cham, Switzerland: Springer, 2023, pp. 93–106.
[20] I. D. Mienye and N. Jere, ‘‘A survey of decision trees: Concepts, algorithms, and applications,’’ IEEE Access, vol. 12, pp. 86716–86727, 2024.
[21] L. Wang, ‘‘Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization,’’ Appl. Soft Comput., vol. 114, Jan. 2022, Art. no. 108153.
[22] I. Emmanuel, Y. Sun, and Z. Wang, ‘‘A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method,’’ J. Big Data, vol. 11, no. 1, p. 23, Feb. 2024.
[23] A. F. Aysan, B. S. Ciftler, and I. M. Unal, ‘‘Predictive power of random forests in analyzing risk management in Islamic banking,’’ J. Risk Financial Manage., vol. 17, no. 3, p. 104, Mar. 2024.
[24] J. Liu, J. Liu, C. Wu, and S. Wang, ‘‘Enhancing credit risk prediction based on ensemble tree-based feature transformation and logistic regression,’’ J. Forecasting, vol. 43, no. 2, pp. 429–455, Mar. 2024.
[25] Singh, M., Tiwari, S. K., Swapna, G., Verma, K., Prasad, V., Patidar, V., Sharma, D. & Mewada, H. (2023). A Drug-Target Interaction Prediction Based on Supervised Probabilistic Classification. Journal of Computer Science, 19(10), 1203-1211. https://doi.org/10.3844/jcssp.2023.1203.1211
[26] M. Mahbobi, S. Kimiagari, and M. Vasudevan, ‘‘Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural

networks,'' Ann. Operations Res., vol. 330, nos. 1–2, pp. 609–637, Nov. 2023.

[27] A. Nazemi and F. J. Fabozzi, ''Interpretable machine learning for creditor recovery rates,'' J. Banking Finance, vol. 164, Jul. 2024, Art. no. 107187.

[28] F. M. Talaat, A. Aljadani, M. Badawy, and M. Elhosseini, ''Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction,'' Neural Comput. Appl., vol. 36, no. 9, pp. 4847–4865, Mar. 2024.

[29] P. J. G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann, ''The coming of age of interpretable and explainable machine learning models,'' Neurocomputing, vol. 535, pp. 25–39, May 2023.

[30] Ganesh, B. R., B M, P., Prasad K, K., Swapna, G., & G, Viswanath. (2025). Data Mining-Driven Multi-Feature Selection for Chronic Disease Forecasting. Journal of Neonatal Surgery, 14(5S), 108–124.