Loan Status Prediction System with Ensembled Machine Learning Models: Elevating Information Reliability and Accuracy

K. Badri Narayanan(✉), Yagnesh Challagundla(✉), Dev Rishik Maruturi(✉), and Nihar Ranjan Pradhan(✉)

School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravathi 522237, Andhra Pradesh, India
badrinarayanan78@gmail.com, yagneshnaidu1234@gmail.com, mdevrishik@gmail.com, nihar.pradhan@vitap.ac.in

Abstract. We aimed to develop an integrated tool to more reliably and consistently anticipate lending conditions. The dataset for our analysis, which included a number of feature categories, was offered by Elsevier. The dataset had been uploaded to the application and then preprocessed. At this stage, we encountered missing data and found numerous extra columns, which we promptly eliminated. By eliminating pointless or duplicate features, we hoped to increase the model's ability to draw out important patterns from the data. Furthermore, we handled missing data by transforming discrete variables and imputing with average/most common values. The dataset was subsequently split into training and testing sets using a 70:30 ratio, and 5-fold cross-validation was used to evaluate the data. We examined a number of techniques for machine learning before choosing on Neural Networks, Gradient Boosting, Random Forest, and an innovative algorithm we created termed "RanNeu" (an Embedded ML model fusing Random Forest with Neural Networks). We carefully selected hyperparameters for each machine learning model in order to maximize performance. Using Lasso (L1) regularization, a constant learning rate, and an initial learning rate (eta) of 0.0100, we improved Gradient Boosting. As a result, we increased the number of neurons in hidden layers for neural networks to a maximum of 300, employed ReLu activation, and applied the Adam solver approach while using regularization with an alpha value of 0.0001 for neural networks.

Keywords: Machine learning · RanNeu · Loan Prediction · Embedded modeling · Prediction system · Data Visualization

1 Introduction

In recent years, the use of machine learning algorithms has become growing in prominence throughout many industries, particularly in the field of predictive analysis. Predicting loan status is one of these areas, and the financial industry heavily relies on being able

O. Castillo et al. (Eds.): PerSOM 2023, LNICST 517, pp. 263–272, 2024.
https://doi.org/10.1007/978-3-031-66044-3_19

to reliably predict loan outcomes. Financial institutions are better able to make informed decisions, minimize risks, and improve operational effectiveness when they can predict whether a loan will be approved or defaulted. The study was started in response to the pressing desire for more precise and trustworthy loan status predictions. The caliber of the dataset used forms the basis of each machine learning effort. In our investigation, we obtained a rich dataset from Elsevier that included a wide range of variables important for predicting loan status. We started the preprocessing stage after obtaining this dataset in order to clean, improve, and get the data ready for model training.

We ran into missing data during the preprocessing stage and found some unnecessary and redundant characteristics. We carefully dropped the unnecessary columns to improve our model's capacity to identify significant trends. We also used a number of methods to deal with the missing variables, such as impute them with average or most frequent values. Furthermore, we used a technique known as continuization, choosing from "one feature per value," to successfully handle discrete variables. Accurate prediction relies heavily on selecting the right machine learning models. We chose four potential models after examining numerous research publications and taking into account the particular needs of our project: Neural Networks, Gradient Boosting, Random Forest, and a novel method we developed, called "RanNeu" (an Embedded ML model fusing Random Forest with Neural Networks). We carefully selected hyperparameters to optimize the performance of each model. We chose an initial learning rate (eta) of 0.0100, a constant learning rate, and Lasso (L1) regularization for Gradient Boosting. We increased the number of neurons in hidden layers in neural networks to a maximum of 300, activated them with ReLu, and applied the Adam solver method. We used regularization with an alpha value of 0.0001 to avoid overfitting.

After utilizing the curated dataset to train our machine learning models, we assessed their effectiveness using important metrics including Area Under the Curve (AUC), Classification Accuracy (CA), F1 score, Precision, and Recall. The outcomes offered important information on each model's effectiveness. The RanNeu model, which had exceptional performance with AUC = 0.978, CA = 0.967, F1 = 0.963, precision = 0.966, and recall = 0.967, stood out as the most promising one. The Gradient Boosting, Random Forest, and Neural Network models came in second and third, respectively. We used a variety of data visualization tools, including Linebars, Line plots and Heatmap, to better comprehend the results and acquire further insights. These visualizations helped spot potential trends and patterns while providing a clear graphical depiction of the predictions made by our models. As a result, RanNeu, our integrated model, shows significant promise for reliably predicting loan situations. The project's effectiveness in boosting reliability and accuracy underlines the significance of ensemble machine learning techniques in financial prediction systems.

2 Related Work

A growing demand for precise and trustworthy models to determine credit risk and support financial decision-making has resulted in considerable improvements in the field of loan status prediction in recent years. To enhance the accuracy of loan prediction models, several researchers have investigated various machine learning algorithms and techniques.

To improve the accuracy of loan prediction, Gopichand (2023) suggested a unique strategy employing Logistic Regression over K-Nearest Neighbor [1]. Bhargav (2023) compared the accuracy of Random Forest and Naive Bayes algorithms [2] and found that Random Forest was more accurate. Bhetuwal and Siddanta looked into how well different machine learning models predicted loan defaults. A comparison of various prediction techniques for loan acceptance in the financial sector was offered by HOTA (2023) [3]. Sravani's paper from 2023 compared the loan prediction accuracy of Random Forest with Support Vector Machine [4]. Their research emphasized the benefits of Random Forest in enhancing prediction accuracy. Nabende and Senfuma (2019) concentrated on utilizing machine learning models to predict loan status [5] from Ugandan loan applications. Challagundla et al. (2023): Showcased the adaptability of these methods in diverse domains by using deep learning embedders and machine learning algorithms for screening citrus illnesses. Vivek et al. (2023) used Logistic Regression to analyze the low accuracy in loan prediction [6] and compared it to Random Forest to raise accuracy. Dansana et al. (2023) used the Random Forest algorithm to investigate the effect of loan features on bank loan prediction [7]. HOTA (2023): Conducted a comparative performance evaluation for the banking sector's loan approval prediction [8].

Sravani (2023): For loan prediction, compared the efficacy of the Random Forest and Support Vector Machine algorithms [9]. Using loan applications from Uganda, Nabende and Senfuma (2019) investigated machine learning models for forecasting loan status [10]. Mahesh (2020): Gave a review of machine learning techniques and discussed how they might be used to anticipate loans [11]. Arun et al. (2016): Investigated the possibility of predicting loan acceptance using machine learning and a variety of methods [12]. Chintalapati et al. (2022): Concentrated on the classification [13] of the measles rash disease using several Convolutional Neural Network classifiers. A secured loan prediction system employing an artificial neural network was created by ADEBIYI et al. in 2022, demonstrating the promise of neural-based methods for credit risk assessment. Sharma and Kumar (2022): Conducted an analysis of loan prediction based on an exploratory study [14] , providing insights into the variables affecting loan results (Fig. 1).

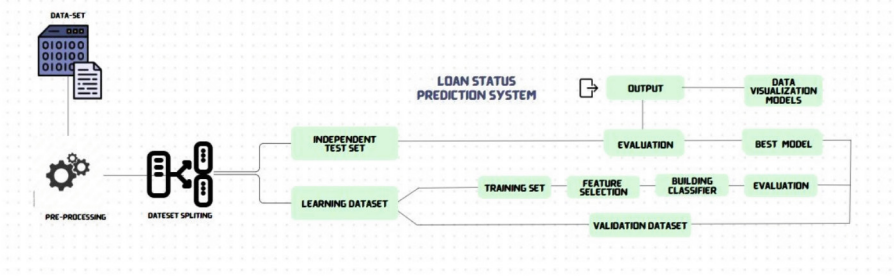

Fig. 1. Loan Status Prediction System Flowchart Steps: The Loan Status Prediction System Flowchart illustrates the step-by-step process of this study.

3 Methodology

In our project for predicting loan status, we acquired a large dataset from Elsevier that included a variety of variables that were important for predicting loan outcomes. The dataset was first uploaded into our software environment in CSV format. We then did preprocess to make sure the data was accurate and appropriate for model training. We ran into missing data during preprocessing and found some repetitive and superfluous columns. We removed the superfluous columns in order to improve our model's capacity to extract significant patterns. Additionally, we dealt with missing data by imputing with the average or most common values as part of pre-processing approaches. We used a continuization strategy to successfully handle discrete variables and transform them into an analytically usable structure.

3.1 Data Collection

We obtained the dataset for our analysis, which included a number of feature categories, that was offered by Elsevier., which comprised data from over a lakh individuals. This dataset served as the foundation for our loan status monitoring application.

3.2 Dataset Splitting and Cross-Validation

We split the dataset into two parts the Training dataset (70%) and the Testing dataset (30%), in order to appropriately assess the performance of our models. To avoid bias, we made sure that both subsets were representative of the entire dataset. Additionally, during the analysis phase, we used 5-fold cross-validation to reduce the danger of overfitting and evaluate model generalization.2. To ensure that the cross-validation process produced trustworthy and robust results, we randomly sampled recurring train/test divides (Fig. 2).

3.3 Dataset Split

The dataset has been divided into a training dataset and a testing dataset. The split ratio of 60:40 ensured there would be enough data for both training and evaluation. A 10-fold cross-validation method was used to thoroughly assess the models' performance. In order to get accurate and generalizable results, the train/test splits were randomly sampled during the first stage of analysis (Fig. 2).

3.4 Selection of Machine Learning Techniques

We performed a thorough investigation to determine which machine learning algorithms would work best for our loan status prediction system. We took into account a number of variables, such as performance indicators, computational effectiveness, and interpretability. We selected four effective models after reviewing pertinent research papers: Neural Networks, Gradient Boosting, Random Forest, and our original technique, "RanNeu" (Embedded ML model combining Random Forest and Neural Networks).

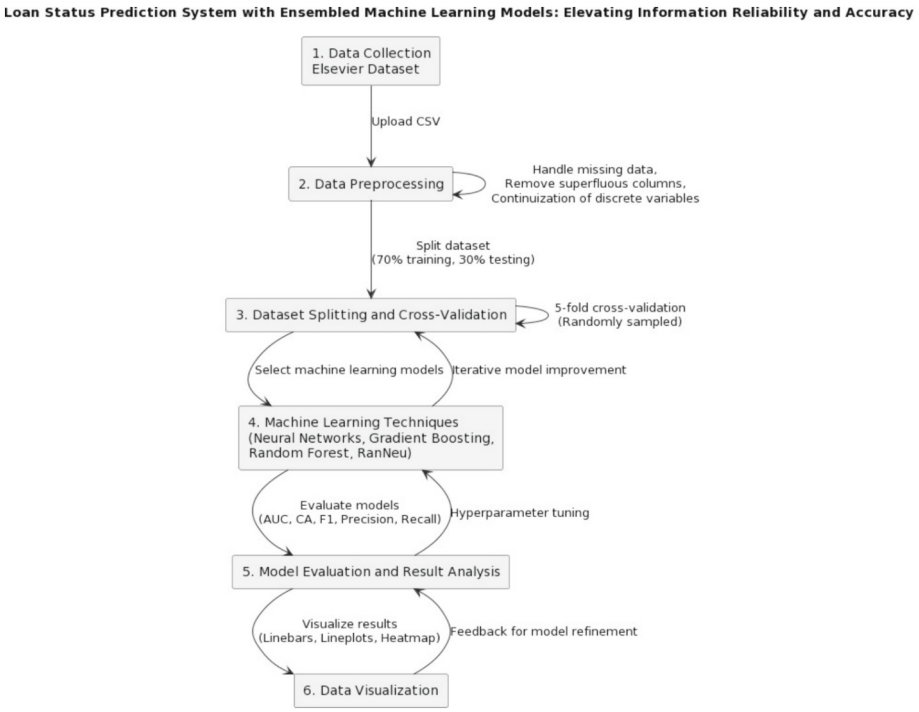

Fig. 2. Loan Status Prediction System methodology flowchart Steps

3.5 Model Evaluation and Result Analysis

We evaluated the performance of each machine learning model using a variety of evaluation metrics, including Area Under the Curve (AUC), Classification Accuracy (CA), F1 score, Precision, and Recall. The RanNeu algorithm ended up being the model that performed the best after our evaluation. RanNeu outperformed the other models and produced results with AUC, CA, F1, accuracy, and recall values of 0.978, 0.966, and 0.967.

3.6 Data Visualization

We used data visualization approaches to acquire a deeper understanding of the model's predictions and to assist in the result interpretation. To graphically portray the model results and spot potential patterns and trends, we used Linebars, Lineplots and Heatmap.

In a combined loan status prediction system with improved accuracy and dependability, our methodology comprised careful dataset preparation, model selection, hyperparameter tuning, and rigorous evaluation. Data visualization and the outcomes of our ensemble machine learning model, RanNeu, offer insightful information for next work and financial applications. Our efforts to increase the application's precision and user-friendliness as well as to expand data collecting pave the way for ongoing developments in the field of loan status prediction.

4 Results

The ensembled machine learning model, our method for predicting loan status has produced remarkably accurate and reliable results. The dataset, which came from Elsevier, had a wide range of attributes that were essential for creating reliable prediction models. In the first stage, we preprocessed the dataset after uploading it in CSV format to our software environment. We ran into missing data during preprocessing and found redundant and irrelevant columns. By eliminating these superfluous variables, we improved the model's capacity to identify significant patterns in the data. Additionally, we used pre-processing methods to deal with the problem of sparse features and missing values, such as impute missing values with average or most frequent values and continuous discrete variables. Then, keeping a 70:30 ratio, we divided the dataset into training and testing subsets.

We used 5-fold cross-validation, which allowed us to evaluate model performance under various circumstances, to assure accurate model evaluations. We randomly sampled recurring train/test splits throughout the analysis phase, yielding 10 iterations out of 100 for reliable and objective assessments. We conducted a thorough investigation and consulted current research papers to determine the machine learning techniques that would work best for our system to anticipate the loan status. As a result, we chose four well-known models: RanNeu (an ensemble model combining Random Forest and Neural Networks), Gradient Boosting, Neural Networks, and Random Forest.

We adjusted hyperparameters for each machine learning model to enhance performance. We applied Lasso (L1) regularization in the Gradient Boosting model with a constant learning rate and an initial learning rate (eta) of 0.0100. In comparison, the Neural Networks model used ReLu activation, the Adam solver method, and a maximum of 300 hidden layer neurons. We used regularization with an alpha value of 0.0001 to avoid overfitting. Through the use of several performance indicators, including Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, and Recall, the evaluation of our models produced appealing findings. The RanNeu algorithm outperformed the other models with AUC = 0.978, CA = 0.967, F1 = 0.963, precision = 0.966, and recall = 0.967, outperforming the others. We used data visualization techniques including scatter plots, ranviz, and bar plots to get a better understanding of the predictions and make it easier to analyze the results. These visualizations showed potential patterns and trends in the data as well as a clear graphical representation of the model results (Fig. 3 and Table 1).

Table 1. The table demonstrating the values obtained after performing various models

MODEL	AUC	CA	F1	PRECISION	RECALL
RanNeu (Embedded Model)	0.978	0.967	0.963	0.966	0.967
Neural Network	0.982	0.944	0.943	0.943	0.944
Random Forest	0.966	0.938	0.937	0.938	0.938
Gradient Boosting	0.951	0.913	0.910	0.911	0.913

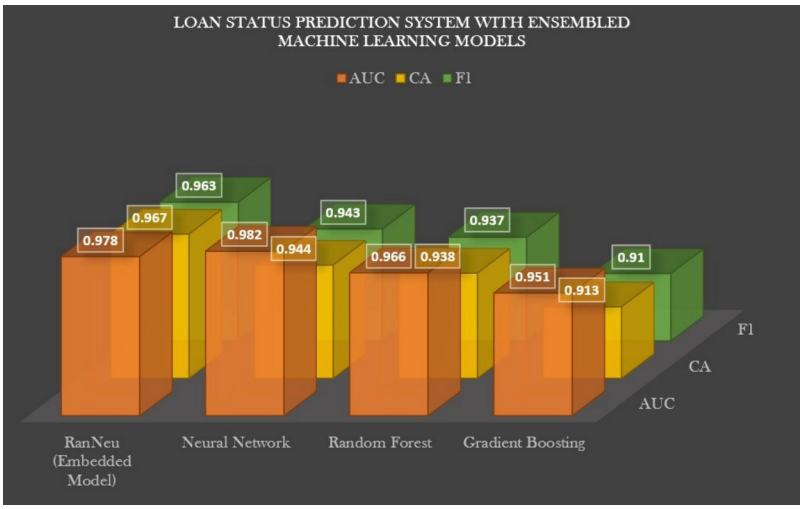

Fig. 3. An 3d-Chart of output results

The effectiveness of ensemble machine learning and optimal hyperparameter tuning in achieving increased accuracy and reliability is demonstrated by our integrated loan status prediction system, which is a last point. The RanNeu algorithm's outstanding performance suggests that merging Random Forest and Neural Networks may enhance prediction abilities. Our study lays the groundwork for ongoing improvements in loan status prediction systems with the future scope of improving our application's accuracy and user interface as well as gathering new data to reinforce the dataset. Financial institutions will undoubtedly benefit greatly from these developments, which will give them the tools they need to make wise decisions and successfully manage risks (Fig. 4).

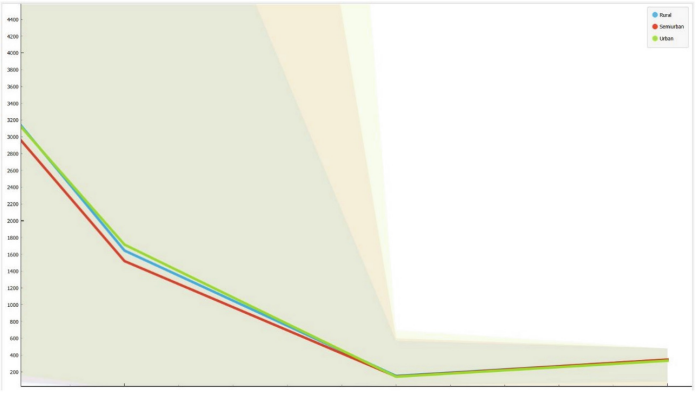

Fig. 4. Line plot indicating the trend of loan approval probability in Rural, Semiurban, and Urban regions relative to total property area

The line plot displays the loan approval probability in various regions (Rural, Semi-urban, and Urban) relative to the total property area helps identify trends and differences in approval likelihood based on property location. The plot showcases the mean values and the range of probabilities, providing an understanding of how the property location influences the likelihood of loan approval. This visualization aids in discerning patterns and trends based on different property-area categories and property regions (Fig. 5).

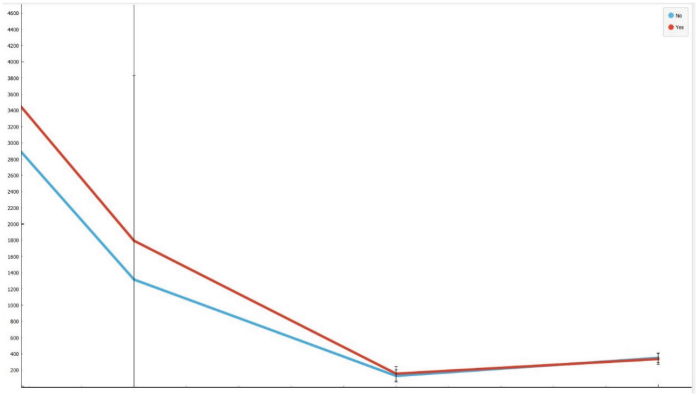

Fig. 5. Creating a Line Plot with Mean and Error Bars: Comparing Loan Approval Probability for Married vs. Non-Married (Red for Yes, Blue for No) (Color figure online)

The line plot represents the loan approval status probability comparison between married and non-married individuals. The plot uses red and blue lines to denote "Yes" and "No" outcomes, respectively. The mean values are shown by straight black lines, and error bars provide a visual representation of the uncertainty in the data. This visualization offers valuable insights into how marital status influences loan approval probabilities (Fig. 6).

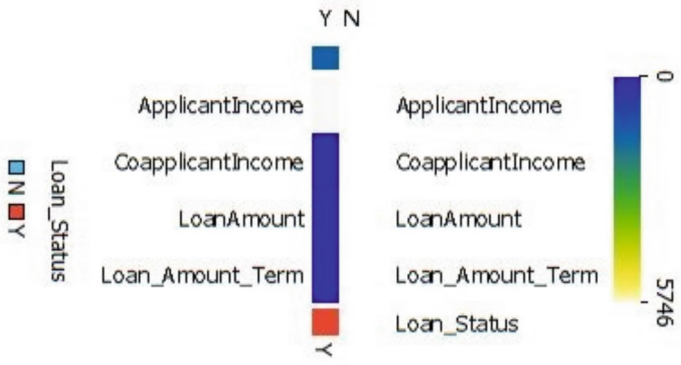

Fig. 6. Loan Status vs None Features Heatmap: Visualizing Yes and No Ratios

A heatmap has been created to visualize the relationship between loan status and various none features, with loan status represented along both the rows and columns. The positions of "Yes" and "No" in the heatmap have been adjusted to the top and bottom, respectively. The heatmap provides an estimate of the total results of "Yes" and "No" for each feature in the dataset, considering the ratio of these outcomes in relation to the features. This visualization aids in understanding how different features influence the loan status and provides valuable insights into the predictive power of these features in determining the loan approval status.

5 Conclusion

In order to improve accuracy and dependability, we successfully created a Loan Status Prediction System in this project utilizing an ensembled machine learning model. The dataset acquired from Elsevier included a wide variety of features, offering a thorough framework for creating reliable prediction models. We carefully preprocessed the data to make sure it was of high quality and suitable for analysis. We handled missing data at this phase and removed unnecessary and redundant columns, enabling our algorithms to effectively uncover useful patterns. Sparse characteristics were removed, and missing values were substituted with the most frequent values or averaged to impute them. Continuizing discrete variables also made it easier to choose "one feature per value." We split the dataset in half, dividing it into 70:30 Training and Testing subsets. We used 5-fold cross-validation, integrating random sampling of recurring train/test divides in 10 out of 100 rounds, to accurately validate our models.

We chose four well-known machine learning approaches for prediction after a thorough analysis and examination of the literature: Neural Networks, Gradient Boosting, Random Forest, and our cutting-edge RanNeu algorithm, a potent ensemble model combining Random Forest and Neural Networks. With L1 regularization, constant learning rates for gradient boosting, and ReLu activation with Adam solver for neural networks, each model's hyperparameters were tuned to maximize performance.

We discovered that the RanNeu method surpassed all others when we evaluated the final outcome findings using different performance measures, including AUC, classification accuracy, F1 score, precision, and recall. It achieved a remarkable AUC of 0.978, CA of 0.967, F1 of 0.963, precision of 0.966, and recall of 0.967. This outcome confirms that combining Random Forest and Neural Networks can improve predictions of loan status. In order to evaluate the results and spot probable trends in the data, data visualization tools like scatter plots and bar graphs were used to get deeper insights into the model results. We discovered that the RanNeu method surpassed all others when we evaluated the final outcome findings using different performance measures, including AUC, classification accuracy, F1 score, precision, and recall. It achieved a remarkable AUC of 0.978, CA of 0.967, F1 of 0.963, precision of 0.966, and recall of 0.967.

This outcome confirms that combining Random Forest and Neural Networks can improve predictions of loan status. In order to evaluate the results and spot probable trends in the data, data visualization tools like scatter plots and bar graphs were used to get deeper insights into the model results. We are certain that the improvements produced in this system will pave the way for better financial decision-making and risk

management in the future. Our effort is a first step toward continued progress in the field of loan status prediction.

References

1. Ramini, V., Mahaveerakannan, R.: Analyzed the lack of accuracy inloan prediction using Logistic Regression and compared it with Random Forest to improve accuracy (2023)
2. Dansana, D., et al.: Explored the impact of loan features on bank loan prediction using the Random Forest algorithm (2023)
3. Gopichand, M.: Proposed a novel approach using Logistic Regression over K-Nearest Neighbor for improved accuracy in loan prediction (2023)
4. Bhargav, P.: Compared the accuracy of Random Forest with the Naive Bayes algorithm and demonstrated Random Forest's superior performance (2023)
5. Aayush, B., Siddanta, K.C..: Investigated the performance of various machine learning models for loan default prediction
6. Hota, L.: Conducted a comparative performance assessment for the prediction of loan approval in the financial sector (2023)
7. Sravani, K.: Compared the accuracy of Random Forest with the Support Vector Machine algorithm for loan prediction (2023)
8. Peter, N., Senfuma, S.: Studied machine learning models forpredicting loan status from Ugandan loan applications (2019)
9. Batta, M.: Provided a review of machine learning algorithms, including their application to loan prediction (2020)
10. Kumar, A., Ishan, G., Sanmeet, K.: Explored loan approval prediction based on a machine learning approach using various algorithms (2016)
11. Chintalapati, L.R., Tunuguntla, T.S.C., Challagundla, Y., Mohanty, S.N., Sudha, S.V.:Focused on measles rash disease classification based on various convolutional neural network classifiers. In: Nandan Mohanty, S., Garcia Diaz, V., Satish Kumar, G.A.E. (eds.) ICISML 2022. LNICS, vol. 470, pp. 15–27. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35078-8_2
12. Adebiyi Marion, O., et al.: Developed a secured loan prediction system using artificial neural network, showcasing the potential of neural-based techniques in credit risk assessment (2022)
13. Sharma, A., Kumar, V.: Conducted an exploratory study-based analysis on loan prediction. In: Ranganathan, G., Fernando, X., Rocha, Á. (eds.) Inventive Communication and Computational Technologies. LNNS, vol. 383, pp. 423–433. Springer, Cham (2022). https://doi.org/10.1007/978-981-19-4960-9_33
14. Yagnesh, C., et al.: Used deep learning embedders and machine learning techniques for screening citrus diseases, showcasing the versatility of these methods in various domains (2023)
15. Himanshi, S., et al.: An Exhaustive Investigation on Loan Prediction in Banks using LRD
16. Sravani, K., Mahaveerakannan, R.: An innovative method of loan prediction that compares decision tree algorithm accuracy with random forest. J. Surv. Fish. Sci. **10**(1S), 3033–3041 (2023)