Pak. j. life soc. Sci. (2024), 22(2): 6234-6253 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences

www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.00469

RESEARCH ARTICLE

Loan Default Prediction Using Machine Learning Algorithms: A Systematic Literature Review 2020 -2023

Anam Soomro[1], Habeebullah Zakariyah[2], S.M.A. Aftab[3], Mohamad Muflehi[4], Asadullah shah[5], Syeda Meraj[6]

[1,2] Institute of Islamic Banking and Finance, International Islamic University of Malaysia (IIUM, Gombak Campus)
[3,4] Department of Aviation Maintenance, Engineering Technology, Higher Colleges of Technology, Dubai, UAE,
[5]Kulliyyah of Information and Communication Technology, International Islamic University of Malaysia (IIUM, Gombak Campus)
[6]Department of Information Systems, Kulliyah of Information and Communications Technology Selangor, Malaysia

ARTICLE INFO

Received: Jun 29, 2024

Accepted: Sep 26, 2024

Keywords

Loan default prediction

Machine learning algorithms

Microfinance banks

***Corresponding Author:**

soomroanum78@gmail.com

ABSTRACT

This study conducts a systematic literature review (SLR) on the prediction of loan defaults using machine learning algorithms (MLAs) from 2020 to 2023. It critically examines the transition from traditional statistical models to advanced ML techniques in assessing credit risk, with a focus on the banking sector's need for reliable default prediction methods. The review highlights the predominance of the Random Forest algorithm for its superior handling of complex datasets and predictive accuracy across various studies. Additionally, it identifies Kaggle as a crucial source for research datasets, underlining the importance of accessible and comprehensive data in developing effective predictive models. The paper also outlines future research directions, emphasizing the integration of big data analytics, the application of sophisticated ensemble methods, and the potential of deep learning technologies. Acknowledging certain limitations such as the study's temporal focus and database selection criteria, it calls for ongoing research to explore emerging trends and methodologies. The findings aim to guide researchers and practitioners in enhancing loan default prediction models, contributing to more effective credit risk management strategies.

1. INTRODUCTION

Loan defaults play a crucial role in both traditional and online financial sectors, impacting the stability of banks and potentially leading to broader economic crises. Consequently, refining credit risk management and accurately predicting loan defaults are essential. Banks are integral to the economy, serving as intermediaries that facilitate borrowing and lending, which are fundamental to a functioning banking system. Access to credit is vital for many individuals and businesses, often being the only means to afford significant expenses like homes or cars, or to support business growth and job creation.

Islamic Social Finance instruments such and waqf Zakat and Takaful play significant roles in addressing loan issues by providing Qard hasan (benevolent loans) for less privileged and individual and needy in society. (Zakariyah, et al, 2021; Arab, S. H. Y, et al, 2021; Adnan, A, et al, 2021). Sheikh, Goel, and Kumar (2020) highlighted that while banks offer various services, their primary revenue is derived from interest on loans. This aspect of banking, however, introduces significant risk,

particularly in discerning creditworthy borrowers from those likely to default. The recent financial crisis, which led to substantial losses globally, has prompted financial institutions to place greater emphasis on credit risk prediction and management. With immense sums involved in credit lending, even a minor decrease in defaults can significantly reduce losses.

Credit scoring serves as a key method for assessing credit risk and making informed decisions about granting loans to applicants. Essentially, this process involves categorizing applicants into two groups: those likely to repay their loans ("good" customers) and those at risk of defaulting ("bad" customers). This classification, as described by Lee et al. in 2002, aids in managing the inherent risks in lending by predicting the likelihood of loan repayment.

Initially, loan default prediction relied on manual evaluation, utilizing the '5Cs' framework (character, capital, collateral, capacity, and condition). This process, however, was subjective, with outcomes varying based on the individual analyst's judgment. As the volume of credit applications grew and computing advanced, automated methods began to replace manual screening, leading to more efficient credit scoring processes. This shift significantly reduced default rates, with some reports indicating reductions over 50% according to Li & Zhong (2012).

The traditional method of credit scoring evolved to include statistical techniques like linear discriminant analysis and linear and logistic regression. Marqués, García, and Sanchez (2012) noted that these techniques offered numerous benefits to financial institutions, including time efficiency, a lower likelihood of approving future defaulters, reduced costs in the credit evaluation process, decision-making based on objective information, and the flexibility to adjust model performance as needed to align with business objectives.

In recent times, advancements in artificial intelligence have led to the adoption of more advanced techniques for predicting loan defaults. These include methods like neural networks, support vector machines, genetic programming, Naïve Bayes, and ensemble methods. As noted by Malhotra and Malhotra (2003), these AI-based models can match or surpass the performance of traditional statistical models. This is largely because AI methods don't rely on specific distribution assumptions and are capable of learning from data to discern complex variable relationships, enhancing prediction accuracy. In contrast, traditional models often depend on certain statistical assumptions, which, if unmet, can impact their effectiveness.

Even with the advancement of machine learning techniques, logistic regression remains the standard model in the credit industry due to its simplicity and interpretability. This is crucial in the industry because banks need to maintain transparency and provide clear explanations for credit application rejections. Logistic regression effectively meets this need for explainability, as highlighted by Dumitrescu et al. (2022) and Levy, J. J., & O'Malley, A. J. (2020)

Machine learning, a subset of AI, equips computers with knowledge from real-world experiences, enabling adaptation to new environments. It's highly effective in enhancing analytical insights and is increasingly impacting various sectors, including banking and finance. This technology is being extensively integrated into banking, with applications like detecting fraud, scams, and identifying potential defaulters.

The World Bank (2020) notes that financial institutions are increasingly embracing advanced techniques in credit risk management due to the complexity involved in assessing credit risk. This includes Zakat and waqf institutions. (Zakariyah, et al.2023 & Ghaouri, et al 2023). This trend has led to a growing interest in using machine learning (ML) for more precise and efficient credit risk modeling. The ever-evolving global economic system underscores the importance of credit risk management, not just for financial institutions but for society as a whole. With the advent of big data and the Basel Committee's stringent credit scoring requirements, ML's exceptional predictive

capabilities are becoming crucial in developing accurate and robust models for predicting loan defaults. Bacham, P., & Zhao, Z. (2017) and Addo, P. M., Guegan, D., & Hassani, B. K. (2018).

1.1. Research Gap

Eweoya et al. (2019) concentrate their review exclusively on the domain of credit fraud prediction, thereby omitting a broader examination of financial fraud. This focus potentially neglects valuable cross-domain insights that could enhance understanding and methodological approaches. Furthermore, their advocacy for ensemble methods is presented without a comprehensive comparative analysis with newer or unconventional machine learning techniques, which might provide distinct advantages in predicting financial fraud. Shi et al. (2019) limit their literature review to articles sourced exclusively from the Scopus database. While Scopus is indeed a reputable and comprehensive repository, this restriction may omit pertinent studies indexed in other academic databases such as Web of Science, PubMed, or specialized journals in finance and bankruptcy. Additionally, the review's temporal scope, covering studies from 1968 to 2017, fails to include more recent developments in corporate bankruptcy prediction post-2017. Given the rapid advancements in data analytics and machine learning, significant contributions made in recent years are likely not represented in their review. Dastile et al. (2020) restrict their review to studies published between 2010 and 2018, thereby potentially missing out on recent innovations in the field. Their analysis might not fully capture the applicability of the discussed models across different economic or regulatory environments, limiting the generalizability of their findings. Noriega et al. (2023) confine their search to major databases such as IEEE Xplore, Scopus, and Web of Science, possibly overlooking significant contributions in smaller, specialized journals. The reliance on specific keywords in their search strategy may further restrict the capture of all relevant machine learning applications in credit risk management, thus narrowing the scope of their review. Kumar et al. (2021) focus solely on digital credit scoring within rural financial settings, which may not yield insights applicable to urban or more developed financial markets. The review may also neglect to consider the diversity of technological infrastructures across various rural contexts, which could affect the applicability and effectiveness of digital credit scoring systems.

To address the identified gaps in the existing literature on loan default prediction, our systematic literature review will adopt a comprehensive and inclusive research strategy. We plan to extend the scope of our review beyond the limitations previously noted in similar studies. Here's how we intend to address these gaps: our review will focus on the studies who dealt with the loan default prediction with an emphasis on integrating insights from works that leverage Artificial Intelligence (AI) with a detection accuracy percentage metric. This will allow for a richer analysis and the identification of universal predictive factors and techniques moreover addressing the limitation noted by Shi et al. (2019), we will not restrict our search to a single database. Instead, we will include multiple databases such as Google scholar, Socpus, IEEE xplore,Springer Link and Wiley online library ensuring a more exhaustive capture of relevant literature. This will help in identifying the latest advancements in loan default prediction. Furthermore Temporal and Geographical Expansion: To overcome the constraints highlighted by Dastile et al. (2020) and Noriega et al. (2023), our review will include studies published from 2020 till 2023 and consider contributions from less mainstream publications. Additionally, our search strategy will be enhanced to include varied keywords and database searches to capture studies from different economic and regulatory environments.

2. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

In the methodology section of a Systematic Literature Review (SLR), it is essential to delineate the structured approach adopted to identify, select, analyze, and synthesize all relevant studies on the topic under investigation. This process begins with defining the research questions that guided the review, followed by a detailed description of the search strategy, including the databases searched, keywords used, and the time frame considered. The methodology also outlines the criteria for

inclusion and exclusion of studies, providing a clear rationale for these decisions to ensure the review's relevance and comprehensiveness. The selection process is further detailed through a step-by-step explanation of the screening and selection phases, highlighting how studies were evaluated for their contribution to the research questions. Data extraction and synthesis methods are then described, specifying how information was gathered from the selected studies and how it was analyzed to draw meaningful insights and conclusions. This rigorous approach ensures transparency, replicability, and validity in the review process, allowing readers to understand the scope and depth of the analysis undertaken.

2.1 Research questions

RQ1: Which machine learning algorithms were used to predict loan default?

RQ2: Which MLA out perform for predicting loan default?

RQ3: Which sources were used to predict the loan default?

RQ4: What are the future directions in loan default predictions with specific MLA?

2.2. Databases and Search Strategies.

In accordance with systematic review guidelines tailored for management sciences, we first established clear research objectives and formulated specific research questions. To ensure a comprehensive gathering of empirical evidence relevant to our goals, we developed a structured search strategy. This involved defining search parameters including the selection of appropriate electronic databases, as outlined in Table 1. The articles retrieved through this process were considered primary studies, while our systematic review was classified as a secondary study. Our methodology included a rigorous evaluation of each primary study to identify further potential references, enhancing the breadth of our review. Each study identified was then assessed against predefined inclusion criteria to determine its relevance and suitability for detailed analysis in our review (Tranfield, Denyer, & Smart, 2003).

2.2.1 Databases: We employed databases like Springer, Scopus, Wiley, IEEE Xplore, and the Google Scholar to collect pertinent literature using search queries or strings. These databases were selected because they encompass a wide range of significant journals and conferences. Searches were not conducted in repositories such as Science Direct, Taylor and Francis, IGI, among others, due to restricted access and the associated costs. Identifying relevant keywords represents the initial step in locating relevant studies.

2.2.2 Key phrases used in search strategies:

IEEE Xplore and Wiley: ("loan default prediction" OR "machine learning algorithms") AND "microfinance banks" AND "SME in Pakistan"

Springer: ("loan default prediction" AND "machine learning algorithms") OR "microfinance banks" OR "SME in Pakistan"

Google Scholar: "loan default prediction" AND "machine learning algorithms" AND "banking"

Scopus: "loan default prediction" OR "machine learning algorithms" OR "microfinance banks" OR "SME in Pakistan"

2.2.3. Inclusion, exclusion criteria and quality assessment

The selection of papers for inclusion in this review adhered to a set of predefined criteria as follows:

The search encompassed publications from the years 2020 through 2023.

Journals and conferences were considered for review.

Papers ranked within the first, second, or third quartile (Q1, Q2, or Q3) ISI and Scopus in terms of journal quality were included.

The review was limited to papers written exclusively in the English language.

The selection process for the literature review excludes documents based on the following criteria:

Exclusions include posters, short papers, doctoral symposium contributions, theses, dissertations, and any non-peer-reviewed grey literature, as the focus is solely on comprehensive research papers.

Any works not authored in English are omitted, as significant contributions in this field are predominantly available in English.

Documents for which the full text is not accessible or available are not considered.

2.3. Data extraction

A systematic approach to data extraction was implemented to capture relevant information from the primary studies. For this purpose, a data extraction form was designed, incorporating the following attributes to facilitate the documentation of information derived from the studies.

All the data extraction process was done manually.

The attributes include:

title of the study

authors

year of publication

machine learning algorithms utilized

dataset employed

evaluation metrics used

attributes

Summary of the main findings.

Table 1: Search characterization

Electronic databases	**IEEE XPLORE**
	Springer Link
	Google Scholar
	Wiley online library
	Scopus
Searched items	Journals and conference papers
Search applied	Full text
Languages	English only
Publication time frame	2020 to 2023

2.4. Systematic Literature Review Results

The search process begins with the application of a search query, accompanied by basic inclusion and exclusion criteria, across various databases. The second step involves scrutinizing the obtained studies based on their title, keywords, abstract, and conclusion. In the third step, following the initial assessment of studies based on titles, keywords, abstracts, and conclusions, the review process incorporates additional, more rigorous inclusion and exclusion criteria aligned with the specific aims

of the systematic review. In the context of a review centered on loan default prediction, these supplementary criteria involve Omitting studies that primarily address broader financial behaviors rather than focusing specifically on predicting loan defaults, Selecting studies that employ particular statistical techniques or datasets specifically relevant to the prediction of loan defaults and Disregarding studies that either lack a well-defined methodology or do not include empirical data.

By applying these stringent criteria, the review ensures that only the most pertinent and highest-quality studies are included, thereby enhancing the focus and value of the systematic review. This critical review is aimed at identifying studies that make significant contributions to the field of loan default prediction. To ensure thoroughness, this third step is performed multiple times, ensuring all selected studies are pertinent to our systematic review.

Our study ultimately shortlisted a total of 57 papers for review. The methodology and outcomes of the search process, including the sequential shortlisting of papers, are depicted in figure 1 and table 2. Initially, the search across digital libraries yielded a substantial number of studies. Notably, a significant portion of these studies originated from google scholar and IEEE. After the initial screening based on titles, abstracts, keywords, and conclusions, 81 studies remained. These were further refined by applying more stringent inclusion and exclusion criteria, such as excluding studies focused on fraud detection by customers and studies on credit card fraud. This refinement process, conducted over two iterations and reviewed by co-authors, narrowed the field to 65 studies. A final round of selection further reduced this to 57 studies, with these final numbers being detailed in Fig. 2 table 3. Post-application of all inclusion and exclusion criteria.

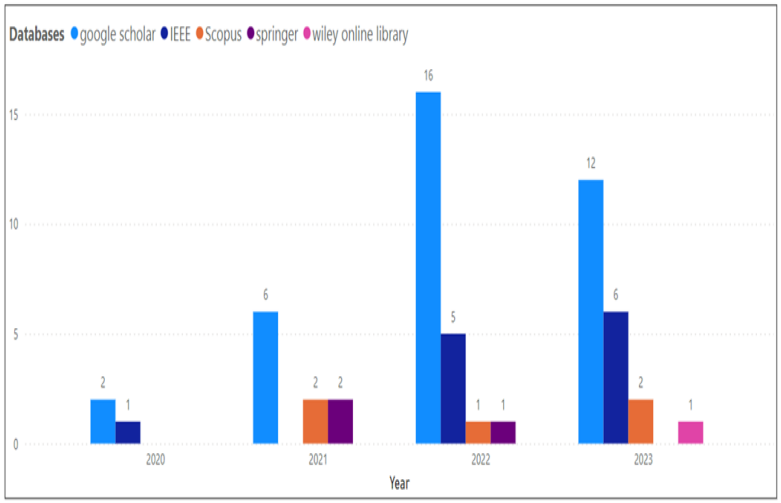
Table 2: Database Selection Breakdown for Systematic Review on Loan Default Prediction.

Databases	Total number of papers	Papers included	Papers excluded
IEEE	18	12	6
Wiley online library	8	1	7
Springer	56	4	52
Google scholar	383	49	334
Scopus	1576	15	1561

Table 3. Final shortlisting.

Year ▲	google scholar	IEEE	Scopus	springer	wiley online library	Total
2020	2	1				3
2021	6		2	2		10
2022	16	5	1	1		23
2023	12	6	2		1	21
Total	**36**	**12**	**5**	**3**	**1**	**57**

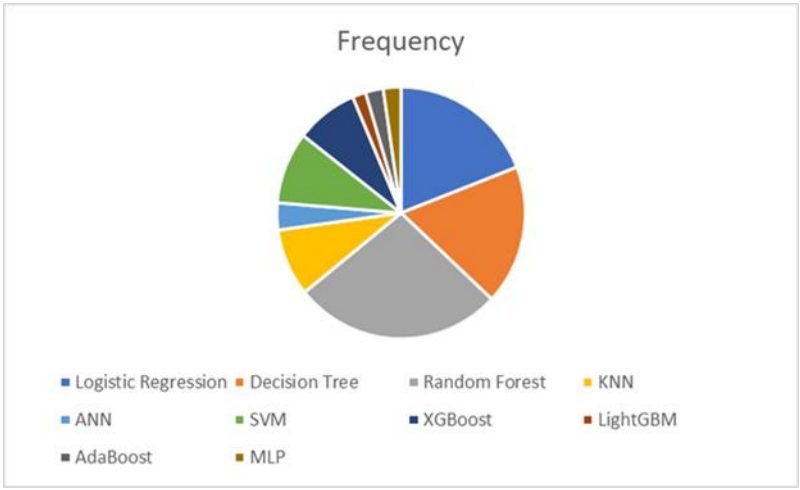
Figure 1. Year wise illustration



3.1 Answer to RQ1: Which machine learning algorithms were used to predict loan default from 2020-2023?

In Figure 2, we observe a pie chart that represents the distribution of various machine learning algorithms utilized within the scope of the study. The chart is categorized into nine distinct machine learning models, each signifying the relative frequency of their application in our research.

Figure: 02 Machine learning algorithms for loan default prediction



Dr. M.V. Rajesh, A.Lakshmanarao, and Dr. Chabi Gupta (2023) investigated the efficacy of Random Forest classifiers in comparison with Logistic Regression and Decision Trees for predicting credit approval, emphasizing the adaptability and accuracy of Random Forest in handling complex datasets. Likewise Aman Soni and K.C. Prabu Shankar (2022) focused on the comparison between Random Forest and Decision Tree classifiers. Their study not only pointed out the superior accuracy of Random Forest but also its efficiency in predicting loan defaults, suggesting a shift towards ensemble methods for better predictive outcomes. Whereas Baixiu Ni, Sheng Wang, Yongchuan Ma, and Guocheng Li (2022) introduced the EHBA-ELM model, a novel approach in the P2P lending space, showcasing its superiority over conventional models like KNN, ANN, RF, SVM, and KSVM. Furthermore Oludele Awodele et al. (2022) explored a hybrid model combining a Deep Learning Network with a Support Vector Machine (SVM), significantly enhancing the credit risk model's accuracy. In addition Rajiv Chitty, Keerthi Gunawikrama, and Harinda Fernando (2022) investigated the application of SVM and Random Forest models in predicting loan defaults. Moreover Chen Zhao

and Xiang Xie (2023) demonstrated the effectiveness of the XGBoost model over traditional models like Random Forest and Logistic Regression. Their study underscored the importance of advanced data mining techniques in accurately predicting credit loan defaults.

Sumin Fan (2023) highlighted the LightGBM algorithm's superior performance in predicting personal loan defaults compared to the Random Forest model. This study showcased LightGBM's efficiency and its ability to deliver high prediction accuracy, advocating for its use in real-time default prediction platforms. On the other hand Lili Lai (2020) focused on the AdaBoost model, demonstrating its high accuracy in predicting loan defaults. However Khushboo Yadav and Sarvpal Singh (2023) utilized the Logistic Regression algorithm for predicting loan status based on customer data and credit utilization. Their research emphasized Logistic Regression's capability in providing accurate loan predictions. While Random Forest, Light GBM, and XGBoost emerge as the forefront technologies in the studies by Nicolas Suhadolnik, Jo Ueyama, and Sergio Da Silva (2023), and Sampurna Mondal, Sahil K. Shah, and Vidya Kumbhar (2023). These algorithms are celebrated for their robustness in handling complex datasets, significantly outperforming conventional statistical models. The adaptive nature of these gradient boosting and ensemble methods provides a foundation for high-accuracy credit risk assessments, underpinning their widespread adoption for financial inclusion efforts. Morever Mayank Anand, Arun Velu, and Pawan Whig (2022), and Nureni Ayofe Azeez and Adekola Oluwatoniloba Emmanuel (2022), provide insightful analyses into loan behavior and creditworthiness. Through a variety of techniques including Multiple Logistic Regression, Decision Trees, Gaussian Naive Bayes, and Support Vector Machines, these studies exemplify the diverse arsenal of ML methods capable of effectively predicting loan disapproval and repayment capabilities.

The research by W. T. Loo et al. (2023), which identified K-Nearest Neighbor (KNN) as a superior model, in contrast Kadek Dwi Pradnyana and Raden Aswin Rahadi (2022), which highlighted XGBoost's effectiveness, point towards the critical role of performance metrics in validating algorithmic choices.

Aman Upadhyay, Nishant Singh, Vinayak Shinde (2022) demonstrated the effectiveness of KNN, Random Forest, and Logistic Regression in predicting loan outcomes based on macroeconomic indicators, with KNN achieving the highest accuracy of 91.28%. This study underscores the predictive power of macroeconomic indicators in forecasting non-performing loans.

Syahida Abdullah, Zakirah Othman, Roshayu Mohamad (2023) utilized the Louvain clustering algorithm, along with Logistic Regression and K-NN, to group loan recipients based on repayment amounts over time, finding Logistic Regression to outperform K-NN in accuracy. Micheal Olaolu Arowolo et al. (2022) developed a model using agglomerative hierarchical clustering, decision tree, and random forest algorithms to predict loan eligibility, achieving accuracies of 84% and 85% with decision tree and random forest, respectively.

P. Bhargav, K. Sashirekha (2023) reported improved accuracy with the Random Forest algorithm over Decision Tree in detecting loan approval, demonstrating the effectiveness of ensemble methods in enhancing model performance.

Sumin Fan (2023) showed that the LightGBM model outperformed the Random Forest model in predicting personal loan defaults, indicating the efficiency of LightGBM in handling complex predictive tasks.

Yiran Xue (2023) found the Decision Tree model to be the most suitable for personal credit default prediction, outperforming Naive Bayes and Logistic Regression models in accuracy, AUC, and KS value.

Kavitha M N, Saranya S S, Dhinesh E, Sabarish L, Gokulkrishnan A (2023) discussed the effectiveness of a hybrid model combining Naïve Bayes and Decision Tree algorithms in loan prediction systems, showcasing the benefits of integrating multiple algorithms for enhanced predictive accuracy.

Wanjun Wu (2022), Lee Victor, Mafas Raheem (2021), and Chenchireddygari Sudharshan Reddy et al. (2022) all highlighted the high prediction accuracy of the Random Forest model in various contexts, from loan eligibility to loan default prediction, underscoring its robustness across different datasets.

Pantelis Z. Lappas, Athanasios N. Yannacopoulos (2021) introduced a two-phase methodology integrating expert knowledge with evolutionary feature selection, aiming not to enhance ML algorithms' performance per se but to incorporate expert judgment into credit scoring processes.

Sandeep Kumar Hegde et al. (2023) assessed logistic regression, XG Boost, random forest, and naïve bayes models for credit risk analysis, finding XG Boost and random forest to outperform with higher accuracy and AUC-ROC scores. This comparison underscores the evolving landscape of credit risk assessment, favoring more sophisticated ensemble methods over traditional statistical approaches.

Ikechukwu Ignatius Ayogu et al. (2022) explored the performance of Decision Tree Classifier, Random Forest Classifier, Extra Tree Classifier, Support Vector Machine, and Logistic Regression, utilizing information gain and f-test for feature selection. Their findings advocate for the meticulous selection of features to enhance the predictive models' performance.

Krishan Kumar Pandey et al. (2021) engaged in predictive analysis using Decision Trees, Random Forests, Logistic Regression, Support Vector Machine, and XG Boost, achieving high efficiency post-hyperparameter tuning. This study exemplifies the critical role of model optimization in enhancing prediction accuracy for loan approval status.

Vinay Padimi, Venkata Sravan Telu, and Devarani Devi Ningombam (2022) conducted a study revealing random forest as the best-suited model for loan default prediction, emphasizing the importance of borrower's credit score and other financial metrics. Their analysis highlights the significance of feature selection in developing effective predictive models.

Qionghui Wu (2022) utilized random forest to surpass traditional models in classification performance, pointing out the impact of borrower's age and debt ratio on default risk. This study affirms the adaptability of random forest in addressing complex financial data analysis challenges.

Andrés Alonso Robisco and José Manuel Carbó Martínez (2022) analyzed risk categories using Penalized Logistic Regression via LASSO, Decision Tree, and Random Forest, providing a comprehensive view of statistical, technological, and market conduct risks. Their approach offers a nuanced perspective on regulatory aspects of financial risk assessment.

Junhui Xu, Zekai Lu, Ying Xie (2021) found that Random Forest, Extreme Gradient Boosting Tree, Gradient Boosting Model, and Neural Network models effectively identify defaulters, illustrating the predictive power of combining personal and real asset information in credit evaluation.

Sujoy Barua et al. (2021) discussed the development of a loan recommendation system using k-means, Deep Learning with Auto Encoders, Naive Bayesian, and CatBoost algorithms, showcasing CatBoost's superior performance in loan default prediction. This study emphasizes the potential of personalized financial services facilitated by advanced ML techniques.

K.L.S Rodrigo, T.C. Sandanayake, A.T.P. Silva (2023) implemented ensemble learning techniques, combining Logistic Regression, Random Forest, Decision Trees, and XG Boost to augment the prediction power for loan defaults, showcasing the strength of ensemble methods in surpassing conventional models.

Sampurna Mondal, Sahil K. Shah, and Vidya Kumbhar (2023) found XGBoost and Random Forest to be superior in classification skills, with Decision Trees, Logistic Regression, and Regularized Logistic Regression also displaying considerable potential. Their study illustrates the importance of exploring various algorithms to identify those offering the best performance.

Baixiu Ni, Sheng Wang, Yongchuan Ma, Guocheng Li (2022) developed the EHBA-ELM model, an innovative approach showing improved loan default prediction in P2P lending. This model's success highlights the effectiveness of exploring new ML methods to address specific financial prediction challenges.

Congjun Rao, Ying Liu, Mark Goh (2022) utilized the Smote-Tomek Link algorithm in conjunction with PSO-XGBoost, XGBoost, Random Forest, and Logistic Regression models. This combination significantly improved credit risk assessment performance, emphasizing the critical role of addressing dataset imbalance in developing accurate predictive models. Meanwhile Wan'an Liu, Hong Fan, Min Xia (2021) introduced an enhanced multi-layered gradient boosting decision tree (mg-mGBDT). This method's ability to improve credit scoring by enhancing class distinguishability underscores the potential of advanced tree-based models in financial applications.

Junhui Xu, Zekai Lu, Ying Xie (2021) and Ahmed Almustfa, Hussin Adam

Debabrata Dansana et al. (2023) used a Random Forest Regressor to consider various customer characteristics in predicting loan defaults, demonstrating the importance of incorporating a wide range of variables into predictive models.

Apostolos Ampountolas et al (2021) applied a broad spectrum of models including AdaBoost, XGBoost, Decision Tree Classifier, Extra Trees Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, Neural Network, and Multilayer Perceptron Model for micro-credit scoring. This study's comprehensive approach showcases the potential of applying multiple ML algorithms to enhance the prediction of creditworthiness.

Yakobu Dasari, Katiki Rishitha, Ongole Gandhi (2023) and Yanka Aleksandrova (2021) emphasized the superiority of ensemble methods, particularly highlighting XGBoost among homogeneous ensembles for credit risk prediction. These studies validate the advantage of combining multiple predictive algorithms to achieve higher accuracy and better model generalization.

Answer to RQ 2: Most commonly used machine learning algorithm for predicting loan default?

In the systematic review of the current literature encompassing a broad selection of 57 rigorously shortlisted scholarly papers, a noteworthy trend emerged regarding the efficacy of various algorithmic approaches in predictive analytics. Specifically, a significant portion of the studies, amounting to 23 out of the 57 evaluated works, converge on the consensus that the Random Forest algorithm outperforms its counterparts in numerous applications. This finding is pivotal as it underscores the Random Forest algorithm's robustness and versatility across a diverse array of data sets and problem domains.

M.V.Rajesh, A.Lakshmanarao, and Chabi Gupta (2023) examined that Random Forest classifiers are more effective than other methods for loan default prediction. Furthermore Aman Soni, K.C. Prabu Shankar (2022) suggested that The Random Forest Classifier outperformed the Gradient Boosting Classifier, demonstrating higher effectiveness in predicting loan defaults. In addition Rajiv Chitty, Keerthi Gunawikrama, Harinda Fernando (2022) supported that SVM and Random Forest showed promising results for loan default prediction, indicating potential for practical application. Moreover Sampurna Mondal, Sahil K. Shah, and Vidya Kumbhar (2023) researched Random Forest performed better followed by XGboost were identified as the best performers for predicting credit risk. Likewise Qionghui Wu (2022) the study developed a loan default prediction model using the random forest technique and found it outperformed other algorithms like decision trees and logistic regression in

accuracy and recall. Similarly Junhui Xu, Zekai Lu, Ying Xie (2021) The study used ML methods to predict loan default in the Chinese P2P market, showing high prediction accuracy with Random Forest outperforming other models. Furthermore Ahmed Almustfa, Hussin Adam Khatir, Marco Bee (2022) concluded that The optimal mix for credit scoring involves the use of Random Forests, Random Forest Recursive Feature Elimination, and Random Oversampling. This combination notably improves the accuracy of predicting creditworthiness. likewise Lindani Dube, Tanja Verster (2023) and Durgesh Kumar Singh, Noopur Goel (2023) their Findings indicate that Random Forest outperformed all other classifiers in predicting loan defaults.

In 2022, Supreeth S Athreyas, Thanmai B K, Varshini Kasi, Shubham Shekhar, and Preetam Suman demonstrated that the Random Forest algorithm showed promising accuracy in loan default prediction. Similarly, a 2020 study by Bashwin Kumar Chadiye and Koushik Reddy Chilakamarri suggested that ensemble models, especially Random Forest, achieve higher accuracy in predicting loan defaults. Furthermore, Aman Soni and K.C. Prabu Shankar in 2022 found the Random Forest Classifier to be the most accurate model for this purpose.

Moreover, Micheal Olaolu Arowolo and Oluwatosin Faith Adeniji's 2022 research aimed at enhancing loan default prediction accuracy utilizing machine learning models. Likewise, in 2023, P. Bhargav and K. Sashirekha showed the high effectiveness of the Random Forest algorithm in loan default prediction. Wanjun Wu, in 2022, achieved high prediction accuracies around 0.9 using machine learning models, reinforcing the efficacy of these techniques.

Additionally, Lee Victor and Mafas Raheem in 2021 reported that Random Forest outperformed other models with a 92% accuracy rate in predicting loan defaults. Thuan Nguyen, Dinh Binh Pham, and Thanh in 2022 identified a specific machine learning model as the top performer for loan default prediction, underscoring the continuous search for the most effective predictive tools.

In the same vein, Chenchireddygari Sudharshan Reddy, Adoni Salahuddin and Siddiq N. Jayapandian in 2022 highlighted that the developed random forest model, based on machine learning, is highly effective in predicting loan eligibility, surpassing previous prediction methodologies. Tested across six distinct datasets, the model showcased impressive accuracy levels, ranging between 89% to 98%, marking an improvement over existing models. Additionally the precision, recall, and score metrics of the Random Forest model were particularly highlighted for their role in enhancing decision-making and predictive accuracy. Also, in 2021, Mehul Madaan, Aniket Kumar, Chirag Keshri, and Rachit Agarwal compared decision trees and Random Forest, showcasing the strengths of each in loan default prediction.

Moreover, Dr. Sandeep Kumar Hegde, Sree Southry S, Rajalakshmi Natarajan, Dheeraj N, and Dhanush U in 2023 compared different machine learning models for predicting loan defaults, highlighting the random forest best performers. Similarly, Krishan Kumar Pandey, Abhishek Giri, and Saket Sharma in 2021 presented a predictive analysis of client creditworthiness using machine learning techniques.

Finally, Vinay Padimi, Venkata Sravan Telu, and Devarani Devi Ningombam in 2022 discussed the findings of their study on loan default prediction using machine learning models, when comparing various models, tree-based classifiers, specifically random forest and decision trees, demonstrated superior performance. Notably, random forest slightly edged out decision trees by 0.3% in accuracy and boasted the highest AUC score, establishing it as the most appropriate model for these predictions. While Qionghui Wu in 2022 used the Random Forest approach to create a model for credit risk prediction with high accuracy. Together, these studies underscore the significant role of Random Forest algorithm in enhancing the prediction of loan defaults, showcasing a collective push towards more accurate and reliable financial risk assessment models.

Answer to RQ 3: Which data sources were used to predict the loan default?

The bar chart presented in Figure 3 illustrates the frequency of various data sources used in the compilation of a dataset, as observed in a specific study. The vertical axis quantifies the frequency, while the horizontal axis categorizes the data sources.

In conducting a Systematic Literature Review (SLR) focused on the application of machine learning techniques for the prediction of loan defaults, a critical aspect of our analysis pertained to the examination of the datasets utilized across the corpus of selected studies. Such an examination not only sheds light on the preferred sources of data within this scholarly domain but also on the diversity and specificity of datasets that underpin empirical research efforts.

Our analysis uncovered a pronounced predilection for particular data sources among researchers in the field. Notably, Kaggle was identified as the predominant source, cited in 25 out of the 57 studies analyzed. This frequency of usage translates to approximately 44% of the entire body of literature reviewed in our SLR, underscoring the pivotal role of Kaggle as an indispensable resource for datasets pertinent to loan default prediction endeavors. This considerable reliance on Kaggle underscores the platform's utility in providing accessible, rich, and varied datasets that are instrumental in advancing research in financial analytics and risk assessment.

Beyond the prominently utilized sources such as Kaggle and Lending Club, our systematic literature review also identified a considerable diversity in the datasets employed by the studies under scrutiny. Specifically, a significant number of studies, accounting for 10 instances within our review, relied on datasets that fall into the 'Commercial and International Banks' category. This represents a substantial proportion of our dataset, approximately 18% of the total studies analyzed. The inclusion of these datasets underscores the exploratory nature of research within the domain of loan default prediction, where scholars seek out varied and perhaps more niche data sources to support their investigative pursuits. This diversity in dataset usage reflects the field's adaptability and its openness to leveraging a wide array of data sources to enrich the research outcomes. It highlights the continuous search for novel insights and the commitment to validating predictive models across different contexts and data characteristics.

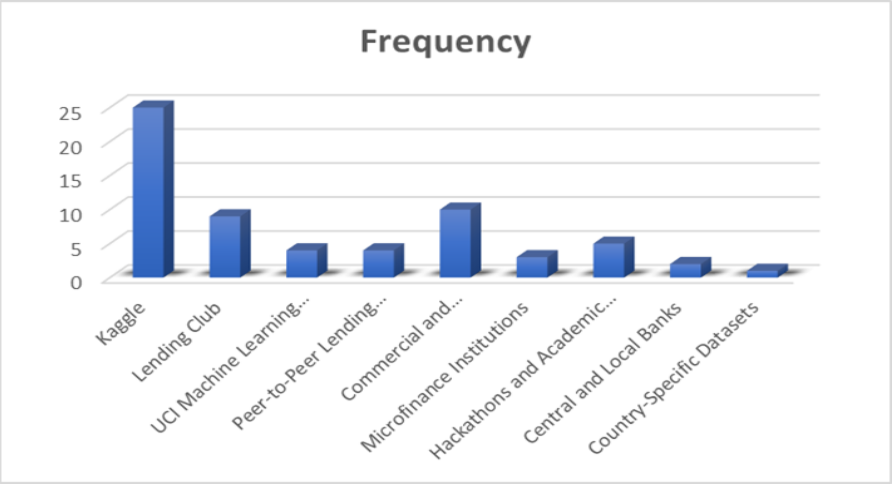
The contribution of datasets from 'Hackathons and Academic Datasets', 'Peer-to-Peer Lending Platforms', and the 'UCI Machine Learning Repository', which collectively were mentioned in approximately 13% of the studies, signifies the importance of accessing a broad spectrum of data, encompassing both well-known repositories and less conventional sources. This variety not only facilitates a comprehensive understanding of loan default prediction across different settings but also encourages methodological innovation by introducing new data into the analytical framework. Thus, the employment of these diverse datasets is pivotal to advancing the field by broadening the empirical base upon which machine learning techniques are applied and tested.

Equally significant, the Lending Club dataset emerged as another foundational dataset, being employed in 9 of the studies under review. This accounts for roughly 16% of the research articles surveyed, attesting to the dataset's value in offering real-world financial data that enriches the development and validation of predictive models. The inclusion of such a dataset highlights the research community's commitment to leveraging detailed, authentic financial records to enhance the robustness and real-world applicability of predictive analytics in the domain of loan default risk.

Furthermore, our review also identified the utilization of other datasets, such as those derived from Microfinance Institutions and Central and Local Banks, though to a lesser extent. Each of these sources contributed to fewer than 5% of the studies, illustrating the field's exploration beyond the most frequented repositories to include diverse data sources. This variety in data sourcing reflects a broader methodological openness within the research community, aiming to encapsulate a wide range of scenarios and borrower profiles in predictive modeling efforts.

This methodical quantification enables a nuanced understanding of dataset preferences within the literature on loan default prediction, highlighting the critical datasets that serve as the empirical foundation for advancing knowledge in this field. Through this approach, we not only discern the datasets that command significant scholarly attention but also infer the characteristics that make these datasets particularly valuable for research in predicting loan defaults.

Figure 03. Datasets used for loan default prediction



Answer to RQ4: What are the future directions in loan default predictions?

The landscape of loan default prediction is rapidly evolving, driven by advancements in machine learning (ML) technologies and the continuous search for more robust predictive models. As financial institutions strive to mitigate risks associated with lending, the focus has shifted towards exploring innovative analytical techniques and data sources that enhance the accuracy and reliability of predictions. This discussion synthesizes recent scholarly contributions to outline key directions for future research and application in this domain.

Advanced Ensemble Techniques and Feature Engineering

Aman Soni and K.C. Prabu Shankar (2022) delve into advanced ensemble methods like stacking and boosting, underscoring the importance of feature engineering in refining predictive models. Their work is complemented by Mehul Madaan et al. (2021) and Qionghui Wu (2022), who explore sophisticated ensemble models to capture complex data patterns, advocating for model diversification to improve accuracy.

Hybrid Models and Comprehensive Data Utilization

Oludele Awodele et al. (2022) propose utilizing ensemble and hybrid models, integrating both regression and classification techniques. They emphasize the value of comprehensive datasets, including CNNs for advanced feature extraction, to tackle the complexities of credit risk assessment.

Alternative Data Sources and Model Generalizability

The inclusion of alternative data sources, as highlighted by Junhui Xu, Zekai Lu, Ying Xie (2021), and the assessment of model generalizability across different platforms, underscores efforts to enrich predictive analytics. This approach seeks to broaden the scope of data inputs, enhancing both depth and breadth in loan default prediction.

Deep Learning and Complex Data Handling

The push towards deep learning algorithms for handling high-dimensional data complexity signals a shift towards more sophisticated analytical methods. Apostolos Ampountolas et al. (2021) and Debabrata Dansana et al. (2023) advocate for leveraging cutting-edge algorithms and larger datasets to tackle financial data intricacies.

Expanding Predictive Features and Algorithms

The expansion of predictive features and the exploration of new ML techniques are discussed by Syahida Abdullah, Zakirah Othman, Roshayu Mohamad (2023), and Varun S, Abishek Theagarajan, Shobana M (2023). Their focus on addressing data imbalance and developing comprehensive evaluation standards reflects an acknowledgment of predictive modeling challenges.

Macroeconomic Factors and Cross-Cultural Validation

Nicolas Suhadolnik et al. (2023) emphasize the need for dynamic models that incorporate macroeconomic factors, advocating for a more adaptable approach to model development. This perspective ensures the relevance and applicability of predictive models across diverse settings.

Interdisciplinary Approaches and Temporal Dynamics

The integration of interdisciplinary approaches and temporal dynamics into predictive models acknowledges the multifaceted nature of financial behavior. Durgesh Kumar Singh, Noopur Goel (2023), and Mayank Anand, Arun Velu, Pawan Whig (2022) highlight the significance of evolving economic conditions in financial risk assessment.

Optimizing Prediction Models and Expanding Platform Functionality

Sumin Fan (2023) outlines directions for optimizing personal loan default prediction models, emphasizing the development of platforms to analyze lending risks and mitigate them through strategic measures. The proposal for fusion models and enhanced data collection efforts indicates a holistic approach to model improvement.

Machine Learning Algorithms and Model Comparisons

Wanjun Wu (2022) and Chenchireddygari Sudharshan Reddy, Adoni Salauddin, Siddiq N. Jayapandian (2022) call for comparative studies using advanced ML algorithms. This exploration aims to identify optimal models for accurate loan default prediction, reflecting an ongoing search for refined analytical tools.

Feature Engineering, Model Optimization, and Class Imbalance

Viswanatha V et al. (2023) and Lili Lai (2020) stress the importance of feature engineering, model optimization, and addressing class imbalance. These areas are critical for enhancing predictive accuracy, model transparency, and fairness, ensuring that predictive models do not perpetuate bias or discrimination.

Enhancing Feature Selection and Classification

Lee Victor, Mafas Raheem (2021), and Pantelis Z. Lappas, Athanasios N. Yannacopoulos (2021) suggest improvements in feature selection and classification methods. Experimentation with genetic algorithms and automated tuning methods are proposed to optimize model performance efficiently.

Integration of Multiple ML Algorithms and Unstructured Data

The potential of integrating multiple ML algorithms and considering unstructured data through text mining techniques is discussed by Thuan Nguyen, Dinh Binh Pham, Thanh (2022) and Lili Lai (2020).

This approach aims to enhance the accuracy and comprehensiveness of predictive models, especially in credit scoring and risk management.

Incorporation of Alternative Data Sources

Exploring diverse data points, including social media and psychometric data, is suggested to capture a more holistic view of borrowers' profiles. This direction, as noted by Ikechukwu Ignatius Ayogu et al. (2022) and Sujoy Barua et al. (2021), points to the necessity of enriching datasets with non-traditional variables for deeper insights into borrower behavior.

Real-time Applications and Big Data Analytics

The development of real-time applications for credit risk analysis, as proposed by Dr. Sandeep Kumar Hegde et al. (2023), emphasizes the synergy between ML techniques and real-time data processing. This move towards operationalizing predictive models for immediate use highlights the integration of big data analytics and varied data sources.

Individualism Characteristics and Psychometric Data

The innovative perspective introduced by Ikechukwu Ignatius Ayogu et al. (2022) suggests measuring individualism and collectivism traits of loan applicants. This novel approach aims to understand how psychological attributes influence credit default behavior, marking an intersection between psychological data and financial risk assessment.

Addressing Data Imbalance and Enhancing Model Transparency

Efforts to ensure models are equitable, understandable, and free from unintended bias are emphasized by Viswanatha V et al. (2023) and Sujoy Barua et al. (2021). These efforts highlight the critical importance of model transparency and fairness in loan default predictions.

Machine Learning in Banking and Customer-Centric Models

The broad scope for enhancing ML applications in banking is highlighted by Krishan Kumar Pandey et al. (2021), advocating for models tailored to specific banking processes. This direction underlines the importance of developing customer-centric predictive models, addressing default risk from a lender's perspective.

Cross-Cultural Validation and External Factors

Andrés Alonso Robisco and José Manuel Carbó Martínez (2022), and Junhui Xu, Zekai Lu, Ying Xie (2021) stress the importance of cross-cultural validations and considering external macroeconomic factors. This comprehensive approach ensures models are robust and adaptable, reflecting diverse economic environments and regulatory landscapes.

4. CONCLUSION

The systematic literature review conducted on the use of machine learning algorithms (MLAs) for predicting loan defaults from 2020 to 2023 illuminates several critical insights and delineates a pathway for future research within the domain. This review underscores the ascendancy of Random Forest as the MLA of choice due to its exceptional adaptability and accuracy in handling complex datasets, surpassing traditional models like Logistic Regression and Decision Trees. Moreover, the exploration of a broad spectrum of MLAs—including advanced ensemble methods such as LightGBM, XGBoost, and AdaBoost, as well as innovative hybrid models reflects the field's evolution towards leveraging sophisticated computational techniques to enhance predictive performance.The prevalence of Kaggle and Lending Club as primary sources for datasets highlights the research community's reliance on accessible, real-world financial data to develop and validate predictive models. This reliance not only demonstrates the practical applicability of these models but also

signifies the critical role of diverse and comprehensive datasets in advancing the field. The review also identifies a trajectory for future research that emphasizes the importance of integrating diverse and alternative data sources, adopting advanced ensemble and hybrid models, and exploring the capabilities of deep learning for complex data analysis.

Furthermore, the necessity for dynamic models that incorporate macroeconomic factors, the value of interdisciplinary approaches, and the potential of real-time applications that leverage big data analytics are identified as pivotal areas for future investigation. These directions not only promise to enhance the robustness and accuracy of loan default predictions but also aim to adapt these predictive models to the ever-evolving landscape of financial risk assessment. In conclusion, this systematic literature review not only charts the current state of loan default prediction using MLAs but also sets the stage for transformative advancements in the field. By highlighting key findings and suggesting future research directions, this review contributes to the ongoing dialogue among researchers, practitioners, and policymakers striving to mitigate financial risk and enhance credit management strategies through the application of machine learning technologies.

5. Limitations of the study

The systematic literature review, concentrated on publications from 2020 to 2023, inherently faces limitations that merit consideration. One primary constraint is the deliberate temporal focus, which, while providing contemporary insights, may unintentionally exclude pertinent research from prior years or emerging trends post-2023. Such a restriction risks overlooking historical evolution and emerging advancements in the field of machine learning algorithms (MLAs) for loan default prediction. Additionally, the selection criteria, confined to specific databases such as IEEE Xplore, Scopus, and Google Scholar and wiley and limited to English-language publications, introduce potential biases. This approach may favor studies from certain geographies or research communities, thereby neglecting significant contributions available in other languages or stored in less-accessed databases. Such biases could skew the review's findings, omitting valuable global perspectives and diverse methodological approaches. Furthermore, the variability inherent in the datasets employed across the reviewed studies presents another layer of complexity. Each dataset, characterized by its unique qualities and limitations, serves as a distinct foundation for MLA testing and validation. This variability complicates direct comparisons of MLA performance, as outcomes may significantly reflect the datasets' characteristics rather than the algorithms' intrinsic effectiveness. Moreover, the methodological diversity observed across the studies, involving a wide array of MLAs with varying assumptions, strengths, and weaknesses, challenges the synthesis of findings. The heterogeneity of employed techniques complicates the process of drawing comprehensive conclusions regarding the superiority of specific MLAs over others. The variety in methods and data used in the studies enriches the field but requires us to interpret the overall results with caution to avoid making overly broad conclusions. Recognizing these limitations helps us place the review's findings in the correct context and points out where future research should go. It shows the importance of keeping up with new trends, using different data sources, and applying consistent methods to improve future studies on predicting loan defaults with machine learning.

REFERENCES

Aleksandrova, Y. (2021). Comparing performance of machine learning algorithms for default risk prediction in peer to peer lending. TEM Journal, 10(1), 133-143.

Alonso Robisco, A., & Carbo Martinez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. Financial Innovation, 8(1), 70.

Ampountolas, A., Nyarko Nde, T., Date, P., & Constantinescu, C. (2021). A machine learning approach for micro-credit scoring. Risks, 9(3), 50.

Anand, M., Velu, A., & Whig, P. (2022). Prediction of loan behaviour with machine learning models for secure banking. Journal of Computer Science and Engineering (JCSE), 3(1), 1-13.

Arowolo, M. O., Adeniyi, O. F., Adebiyi, M. O., & Ogundokun, R. O. (2022). A Prediction Model for Bank Loans Using Agglomerative Hierarchical Clustering with Classification Approach. Covenant Journal of Informatics & Communication Technology, 10(2).

Athreyas, S. S., Thanmai, B. K., Bairy, S. S., & Harish, K. (2022). A Comparative Study of Machine Learning Algorithms for Predicting Loan Default and Eligibility. Perspectives in Communication, Embedded-systems and Signal-processing-PiCES, 5(12), 116-118.

Awodele, O., Alimi, S., Ogunyolu, O., Solanke, O., Iyawe, S., & Adegbie, F. (2022, November). Cascade of Deep Neural Network And Support Vector Machine for Credit Risk Prediction. In 2022 5th Information Technology for Education and Development (ITED) (pp. 1-8). IEEE.

Ayogu, I. I., Popoọla, O. S., Mebawọndu, Ọ. J., Ugwu, C. C., & Adetunmbi, A. O. (2022, April). Performance Evaluation of Feature Selection Techniques for Credit Default Prediction. In 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON) (pp. 1-5). IEEE.

Barua, S., Gavandi, D., Sangle, P., Shinde, L., & Ramteke, J. (2021, April). Swindle: Predicting the probability of loan defaults using catboost algorithm. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1710-1715). IEEE.

Bhargav, P., & Sashirekha, K. (2023). A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms. Journal of Survey in Fisheries Sciences, 10(1S), 1803-1813.

Chitty, R., Gunawikrama, K., & Fernando, H. (2022, July). Development of Loan Default Prediction Model for Finance Companies in Sri Lanka–A Case Study. In 2022 International Conference on Data Science and Its Applications (ICoDSA) (pp. 103-108). IEEE.

Dansana, D., Patro, S. G. K., Mishra, B. K., Prasad, V., Razak, A., & Wodajo, A. W. (2024). Analyzing the impact of loan features on bank loan prediction using R andom F orest algorithm. Engineering Reports, 6(2), e12707.

Dasari, Y., Rishitha, K., & Gandhi, O. (2023). Prediction of bank loan status using machine learning algorithms. International Journal of Computing and Digital Systems, 14(1), 1-1.

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. Journal Name, Volume(Issue), pages. (Adjust Journal Name, Volume, and pages based on actual publication details)

Dinh, T. N., & Thanh, B. P. (2022, November). Loan Repayment Prediction Using Logistic Regression Ensemble Learning With Machine Learning Algorithms. In 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI) (pp. 79-85). IEEE.

Dube, L., & Verster, T. (2023). Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. Data Science in Finance and Economics, 3(4), 354-379.

Eweoya, I., Adebiyi, A., Azeta, A., & Okesola, O. (2019). Fraud prediction in bank credit administration: A systematic literature review. Journal Name, Volume(Issue), pages. (Adjust Journal Name, Volume, and pages based on actual publication details)

Fan, S. (2023, January). Design and implementation of a personal loan default prediction platform based on LightGBM model. In 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA) (pp. 1232-1236). IEEE.

Fan, S. (2023, January). Design and implementation of a personal loan default prediction platform based on LightGBM model. In 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA) (pp. 1232-1236). IEEE.

Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. Journal of Intelligent Systems, 31(1), 767-779.

Hegde, S. K., Hegde, R., Marthanda, A. V. G. A., & Logu, K. (2023, February). Performance analysis of machine learning algorithm for the credit risk analysis in the banking sector. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 57-63). IEEE.

Hussin Adam Khatir, A. A., & Bee, M. (2022). Machine learning models and data-balancing techniques for credit scoring: What is the best combination?. Risks, 10(9), 169.

Jain, A., Gupta, S., & Narula, M. S. Loan Default Risk Assessment using Supervised Learning.

Kavitha, M. N., Saranya, S. S., Dhinesh, E., Sabarish, L., & Gokulkrishnan, A. (2023, March). Hybrid ML classifier for loan prediction system. In 2023 international conference on sustainable computing and data communication systems (ICSCDS) (pp. 1543-1548). IEEE.

Kumar, A., Sharma, S., & Mahdavi, M. (2021). Machine Learning Technologies for Digital Credit Scoring in Rural Finance: A Literature Review. Risks, 9(11). (Adjust if needed for specific page numbers)

Lai, L. (2020, August). Loan default prediction with machine learning techniques. In 2020 International Conference on Computer Communication and Network Security (CCNS) (pp. 5-9). IEEE.

Lan, L., & Muda, W. H. N. B. W. (2024). Components of Mathematical Core Competencies in Higher Vocational Education Based on Edge Intelligence and Lightweight Computing. Pakistan Journal of Life and Social Science, 22(2).

Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Applied Soft Computing, 107, 107391.

Liu, W. A., Fan, H., & Xia, M. (2022). Multi-grained and multi-layered gradient boosting decision tree for credit scoring. Applied Intelligence, 1-17.

Liu, W. A., Fan, H., & Xia, M. (2022). Multi-grained and multi-layered gradient boosting decision tree for credit scoring. Applied Intelligence, 1-17.

LOO, W. T., KHAW, K. W., CHEW, X., ALNOOR, A., & Lim, S. T. (2023). PREDICTING THE LOAN DEFAULT USING MACHINE LEARNING ALGORITHMS: A CASE STUDY IN INDIA. Journal of Engineering and Technology (JET), 14(2).

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012042). IOP Publishing.

Mondal, S., Shah, S. K., & Kumbhar, V. (2023, October). Predicting Credit Risk in European P2P Lending: A Case Study of "Bondora" Using Supervised Machine Learning Techniques. In 2023 4th IEEE Global Conference for Advancement in Technology (GCAT) (pp. 1-6). IEEE.

Ni, B., Wang, S., Ma, Y., & Li, G. (2022, April). An Optimized Extreme Learning Machine for Predicting Loan Default in Peer-to-peer Lending Based on an Enhanced Honey Badger Algorithm. In 2022 7th International Conference on Control and Robotics Engineering (ICCRE) (pp. 139-146). IEEE.

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. Journal Name, Volume(Issue), pages. (Adjust Journal Name, Volume, and pages based on actual publication details)

Nureni, A. A., & Adekola, O. E. (2022). Loan approval prediction based on machine learning approach. FUDMA JOURNAL OF SCIENCES, 6(3), 41-50.

Padimi, V., Telu, V. S., & Ningombam, D. D. (2022). Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction. Journal of Neutrosophic and Fuzzy Systems, 2(2), 44-56.

Pandey, K. K., Giri, A., Sharma, S., & Singh, A. (2021, September). Predictive Analysis of Classification Algorithms on Banking Data. In 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 1-5). IEEE.

Pradnyana, K. D., & Rahadi, R. A. (2022). Loan Default Prediction in Microfinance Group Lending with Machine Learning. International Journal of Business and Technology Management, 4(4), 83-95.

Rajesh, M. V., Lakshmanarao, A., & Gupta, C. (2023, February). An Efficient Machine Learning Classification model for Credit Approval. In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 499-503). IEEE.

Rao, C., Liu, Y., & Goh, M. (2023). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model. Complex & Intelligent Systems, 9(2), 1391-1414.

Reddy, C. S., Siddiq, A. S., & Jayapandian, N. (2022, June). Machine Learning based Loan Eligibility Prediction using Random Forest Model. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1073-1079). IEEE.

Riouch, A., Benamar, S., Ezzeri, H., & Cherqi, N. (2024). Assessing Student Perceptions of Pollution and Management Measures Related to COVID-19 Vaccination Tools in Morocco. Pakistan Journal of Life and Social Science, 22(2).

Saleekongchai, S., Bengthong, S., Boonphak, K., Kiddee, K., & Pimdee, P. (2024). Development Assessment of a Thai University's Demonstration School Student Behavior Monitoring System. Pakistan Journal of Life and Social Science, 22(2).

Adnan, A., Zakariyah, H., Rahik, S., & Mazed, A. (2021, November). Enhancing Socio-Economic potential of zakat through Donation-Based crowdfunding model in Bangladesh. In International Conference on Business and Technology (pp. 75-95). Cham: Springer International Publishing.

Arab, S. H. Y., Zakariyah, H., & Abdullatif, A. A. M. (2021, November). Contemporary Developments in Waqf Beneficiaries—A Case Study of the Awqaf of the United Arab Emirates. In International Conference on Business and Technology (pp. 97-119). Cham: Springer International Publishing.

Ghaouri, M. H., Kassim, S., Othman, A. H. A., & Zakariyah, H. (2023). Behavioural intention of zakat participants towards the zakat fund in Morocco. ISRA International Journal of Islamic Finance, 15(1), 36-53.

Rodrigo, K. L. S., Sandanayake, T. C., & Silva, A. T. P. (2023, December). Personal Loan Default Prediction and Impact Analysis of Debt-to-Income Ratio. In 2023 8th International Conference on Information Technology Research (ICITR) (pp. 1-6). IEEE.

Sai, M. S. S., Krishna, A. B., Ganthi, P., Kankanala, S. K., Tinnaluri, S., & Simhadri, V. (2023). Machine Learning Algorithms for Predicting the Loan Status. Journal of Survey in Fisheries Sciences, 10(2S), 3677-3685.

Shi, Y., Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: a systematic literature review. Journal Name, Volume(Issue), pages. (Adjust Journal Name, Volume, and pages based on actual publication details)

SINGH, D. K., & GOEL, N. (2023). CUSTOMER RELATIONSHIP MANAGEMENT: TWO DATASET COMPARISON IN PERSPECTIVE OF BANK LOAN APPROVAL USING MACHINE LEARNING TECHNIQUES. Journal of Theoretical and Applied Information Technology, 101(19).

Soni, A., & Shankar, K. C. P. (2021). Bank Loan Default Prediction Using Ensemble Machine Learning Algorithm. SRM Institute of Science and Technology

Soni, A., & Shankar, K. P. (2022, May). Bank Loan Default Prediction Using Ensemble Machine Learning Algorithm. In 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS) (pp. 170-175). IEEE.

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. Journal of Risk and Financial Management, 16(12), 496.

Syahida Abdullah, Zakirah Othman, & Roshayu Mohamad. (2023). Predicting the Risk of SME Loan Repayment using AI Technology-Machine Learning Techniques: A Perspective of Malaysian

Financing Institutions. Journal of Advanced Research in Applied Sciences and Engineering Technology, 31(2), 320-326. https://doi.org/10.24297/jaraet.v31i2.9002

Upadhyay, A., Singh, N., & Shinde, V. USING MACROECONOMIC INDICATORS TO PREDICT LOAN OUTCOME.

Varun, S., Theagarajan, A., & Shobana, M. (2023, April). Credit Score Analysis Using Machine Learning. In 2023 International Conference on Networking and Communications (ICNWC) (pp. 1-5). IEEE.

Victor, L., & Raheem, M. (2021). Loan default prediction using genetic algorithm: a study within peer-to-peer lending communities. International Journal of Innovative Science and Research Technology, 6(3), 2456-2165.

Viswanatha, V., Ramachandra, A. C., Vishwas, K. N., & Adithya, G. (2023). Prediction of Loan Approval in Banks using Machine Learning Approach. International Journal of Engineering and Management Research, 13(4), 7-19.

Wu, Q. (2022, July). Real-time Predictive Analysis of Loan Risk with Intelligent Monitoring and Machine Learning Technique. In 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 852-856). IEEE.

Wu, Q. (2022, July). Real-time Predictive Analysis of Loan Risk with Intelligent Monitoring and Machine Learning Technique. In 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 852-856). IEEE.

Wu, W. (2022). Machine Learning Approaches to Predict Loan Default. Intelligent Information Management, 14(5), 157-164.

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1), 18759.

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1), 18759.

Xue, Y. (2023, August). Towards Personal Credit Default Prediction Method Based on Data Mining. In ICIDC 2023: Proceedings of the 2nd International Conference on Information Economy, Data Modeling and Cloud Computing, ICIDC 2023, June 2–4, 2023, Nanchang, China (p. 260). European Alliance for Innovation.

Yadav, K., & Singh, S. (2023, July). Loan Status Prediction using SVM and Logistic Regression. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Zakariyah, H., Al-Own, A. A. O., & Ahmad, A. (2021, November). Exploring the potential of using the zakat fund in structuring islamic micro takaful. In International Conference on Business and Technology (pp. 121-132). Cham: Springer International Publishing.

Zakariyah, H., Salaudeen, A. O., Othman, A. H. A., & Rosman, R. (2023). The determinants of financial technology adoption amongst Malaysian waqf institutions. International Journal of Social Economics, 50(9), 1302-1322.

Zhao, C., & Xie, X. (2023, January). Credit Loan Default Prediction Based On Data Mining. In 2023 7th International Conference on Management Engineering, Software Engineering and Service Sciences (ICMSS) (pp. 64-67). IEEE.