Neural Networks in Credit Scoring: Transforming Financial Assessment

Matteo Montrucchio

November 3, 2024

1 Introduction

A credit score is a prediction of your credit behavior, such as how likely you are to pay a loan back on time, based on information from your credit reports [1].
Financial institutions rely significantly on credit scores to accurately assess whether to grant a mortgage, credit card, or loan. Historically, these scores have been computed using statistical techniques that depend on estimating the probability of default through econometric analysis. However, because of changes in regulations and the increased amount of data, new models utilising machine learning techniques have been created. Machine learning algorithms autonomously recognise the most relevant variables and detect intricate and non-linear relationships between them, which would not have been taken into account in a conventional econometric model.
The Bank of Italy has published a research paper on the utilisation of Artificial Intelligence in credit scoring. The analysis reveals that the new models are generally more accurate in identifying defaults compared to econometric models, with an average improvement that goes from two to ten percent [2]. In recent years, there has been a specific emphasis on utilising neural networks as the primary model for credit scoring assessment.

Hence, the paper will primarily investigate whether the neural network model can serve as an optimal solution for enhancing the prediction of potential defaults and creditworthiness of various entities, including both individual customers and companies. The first paper will examine a neural network designed to determine whether the customer will be able to repay the amount of money that has yet to be paid. The second paper will examine a neural network created by various researchers to evaluate the creditworthiness of customers when applying for a loan from a bank. The third paper will continue to explore the topic of loan authorisation, with a specific emphasis on the peer-to-peer (P2P) lending industry. Next, the fourth paper will explain the application of neural networks in credit scoring for the issuance of credit cards. Lastly, in the fifth paper, there will be an analysis of how a neural network can be implemented to assess the riskiness of decarbonized companies.

Each paper analysis is composed of three distinct components. In the first section, there will be a summary of the researchers' main focus. In the second section, the objective will be on the

[1] Consumer Financial Protection Bureau: the definition of credit score has been taken from the official website of the United States government

[2] Bank of Italy: Research paper *Artificial intelligence in credit scoring: an analysis of some experiences in the Italian financial system* written by Emilia Bonaccorsi di Patti and Massimiliano Affinito

steps taken by the authors to reach to the conclusion that will be discussed in the third section. In the third section, the numerical values deriving from the conclusions will be discussed and used as a parameter of the quality of the paper.

2 State of the Art

2.1 Deep Neural Networks for Behavioral Credit Rating

2.1.1 Summary

Banks' regular course of business includes losses caused by debtors' failure to pay their debts. Predicting a bank's future annual losses is impossible; however, long-term historical data can be used to estimate the average level of credit losses. These estimates, referred to as expected losses, are calculated by multiplying the exposure at default (EAD), loss-given default (LGD), and probability of default (PD). The probability of default can be evaluated using two approaches: behavioural and application models. The loan application process uses application models, whereas behavioural models are used to forecast the future performance of a bank's current credit portfolio. Reliable behavioural models have the potential to serve as an early warning system, allowing banks to offer better loan terms to customers who are at risk of defaulting soon.

Because they believed that deep learning models could capture the complicated relationships between input features and target labels on large datasets, the researchers created a deep forward neural network model for behavioural credit risk assessment. To evaluate the deep forward neural network's reliability, the authors compared it to four other models: two linear (logistic regression and support vector machine) and two non-linear (random forest and a gradient boosting-based model) [Merćep et al., 2021].

2.1.2 Methodology

The research was made possible through a collaboration with a large Croatian bank, which provided a proprietary model development sample for use in this paper. The sample was a behavioural credit risk dataset representing a portion of banks' portfolios between 2009 and 2018. They classified defaulted facilities as positive (or one), non-defaulted ones as negative (or zero), and excluded some facilities from the model. They decided to divide the available data into two separate sets. The first dataset included facilities from 2009 to 2013, while the second dataset covered facilities from 2014 to 2018. This allowed them to compare models trained on data from the global financial crisis to models trained on more recent data. The last year of each dataset was isolated and used to measure model performance, which they called the out-of-time dataset. The remaining years (in-time dataset) were utilised for model development and validation.

The researchers used the following performance metrics:

- Receiver Operating Characteristic curve (ROC curve): it compares the true positive rate (proportion of positives correctly identified as positive) to the false positive rate (proportion of negatives incorrectly identified as positive) [3];

- Area Under the ROC Curve (AUC);

[3] The definition has been taken from the website *Google for Developers*

- Brier score: it measures the accuracy of a probability forecast, the closer is to zero, the better it is [4].

2.1.3 Critical Analysis

For both the datasets (2009 – 2013 and 2014 – 2018), the results showed that all non-linear models achieved better results than linear ones. The random forest model had the worst performance among the non-linear models, although having the best calibrated predictions (indicated by the lowest Brier score). The gradient boosting-based model was very similar with the deep feed-forward network in almost all the performance metrics, except a significant difference between the two models in their Brier scores, with the gradient boosting scoring a much lower 0.039 as opposed to 0.116 for the feed-forward model.

Through this paper, the researchers demonstrated how a deep neural network model for behavioural credit risk assessment can obtain better results compared to linear benchmark model, however, it still does not represent the best possible solution since the results obtained by using a gradient boosting-based model were better. However, in general, the demonstrated difference in performance between linear and non-linear models is significant enough to be beneficial to both banks and clients, so it would make long-term sense to reconsider the regulatory requirements to allow the usage of non-linear models for credit risk assessment purposes.

Unfortunately, the researchers have not discussed in deep the different models but they have simply provided a theoretical description of each one. In addition, in the conclusions, the researchers have just defined the overall performance of the models without mentioning any limitations and possible future work that could be done to improve the performance of the deep feed-forward network.

2.2 Secured Loan Prediction System using Artificial Neural Network

2.2.1 Summary

Lending money has become one of the most complicated tasks for a bank, and to address the issue, financial institutions are increasingly using credit recording procedures to examine loan requests from customers. This is possible because the banking industry has started to collect a large amount of data, mostly from consumer transactions, business financial statements, and payment records. When used and evaluated properly, this data can assist in gaining a competitive advantage and predict consumer creditworthiness, allowing the bank to avoid risky loan disposals.

The most commonly used techniques for loan prediction are logistic regression and discriminant analysis; however, the enormous complexity of the data acquired, due to the non-linear relationship between variables, pushed the researchers to develop a loan prediction system based on an Artificial Neural Network that will determine whether a loan is good or bad, and whether a loan is a payable debt or bad debt. The technology was also designed to anticipate whether a loan application will default on their payments [Adebiyi et al., 2022].

2.2.2 Methodology

The dataset used in the study was derived from customers' historical credit histories at Igboora Micro Finance Bank, and it consisted of 14 columns and 1057 rows.

[4] The definition has been taken from the website *Statistics How To*

The first important phase was the data preparation. Because of the nature of the technique, the researchers converted the dataset's category variables to numerical variables before processing them. The dataset was then divided into two parts: the first was used to train the model, and the second was used to validate the model's results. Finally, they standardised the data to remove the effect of one variable on the others.

In the second phase, they began by setting up the artificial neural network. In the input layer, each customer's details were submitted, and the kernel initializer oversaw the assignment of weights to the various input nodes, even though weight initialisation was quite random at this stage and was usually close to zero. The activation function was applied to the sum of all input nodes with their respective weights to calculate the frequency at which signals would be passed by the neuron. The optimum activation function to use is the one closest to zero, which will be used for the input and hidden layers.

The third phase, the creation of the artificial neural network, was likely the most crucial aspect of the entire procedure. They defined the optimiser, which was the component that assisted in selecting the algorithm capable of determining the most optimal weights for each of the input's nodes.

After fitting the training set of independent and dependent variables to the model, they predicted whether the consumer would repay the loan or not and rated the model based on its accuracy.

2.2.3 Critical Analysis

The model achieved 92% accuracy, showing that the constructed artificial neural network predicts well and may be used to monitor whether a loan applicant would default on his or her repayment or not. The model performed better than linear models, indicating that it can replace models currently employed by financial organisations.

However, one disadvantage of the study is that the criterion used to evaluate the model was just the accuracy, whereas in different research articles on the subject, multiple metrics were examined. This enables for a more accurate comparison of the models and a better understanding of the real potential of the neural network.

The report details with great clarity all of the processes that were taken to create the neural network and demonstrate how the model has a lot of opportunity for development. For example, they suggest that one possible option to improve the model is to make it available to customers on the Internet so that they can see for themselves why their loan would not be accepted based on their characteristics.

2.3 Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction

2.3.1 Summary

Credit card defaults cause problems for both borrowers and financial institutions. Borrowers who fail to pay their credit card obligations may suffer significant financial consequences, including credit score damage, excessive interest rates, and legal action. Credit card defaults can cause considerable costs and reputational harm for financial organisations, as well as regulatory attention and even fines. For this reason, financial institutions utilise a variety of tools and techniques to address credit card defaults, including credit rating models and debt collection companies. As a result, effective credit card default risk prediction models are critical to ensure the financial ecosystem's balance.

The problem with these models is often their interpretability; however, one possible solution is for the researchers to use explainable artificial intelligence (XAI) techniques to develop models that provide accurate predictions, which could help build trust in the models and facilitate their adoption in real-world applications.

The researchers created CreditNetXAI, a revolutionary deep neural network solution for credit card default prediction that combines deep learning and explainable AI techniques. CreditNetXAI has the objective to attain high predicted accuracy while offering results that are easy to understand and transparent [Talaat et al., 2024].

2.3.2 Methodology

The CreditNetXAI method is divided into four basic phases.

- Data Collection and Processing
  The researchers compiled a real-world dataset of 30,000 observations of credit card holders in Taiwan, consisting of 25 variables such as demographic information, credit limits, payment history, and bill amounts. The dataset was imbalanced due to the inclusion of 22120 non-defaulted cases (class 0) and 7880 default cases (class 1). The data processing was easier to perform since the dataset was well-curated and contained no substantial missing data. The outliers were detected using statistical methods like the interquartile range. These processes guaranteed that the dataset used to train and evaluate the model was as representative and reliable as feasible.

- Model Development
  The neural network design was selected due to its ability to capture complex patterns and correlations in credit card default data. The design is made up of numerous layers with proper activation functions for efficiently learning from data. The data was divided into training, validation, and test sets. To avoid overfitting, they trained the model on the training set and monitored the validation loss. Then, they used the validation set to tune hyperparameters such as learning rates, dropout rates, and the number of layers and nodes per layer. Finally, they evaluated the model's performance on the test set. Lastly, they used SHAP to add the XAI component. This is a model-independent explanation approach that calculates feature importance scores and provides insights into the elements that influence the model's predictions.

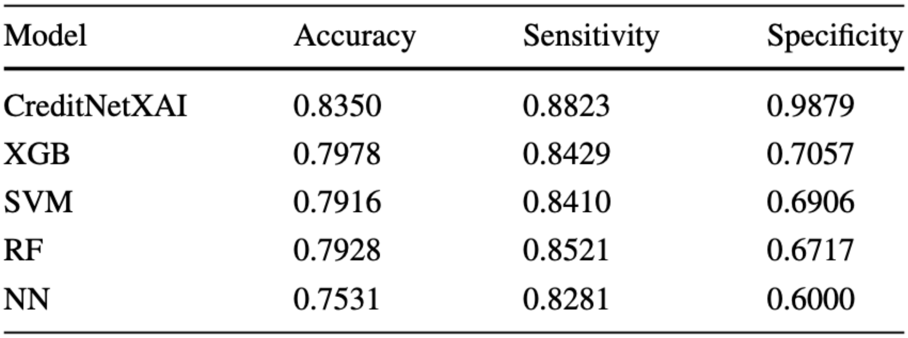
- Model Training and Validation
  The researchers divided the preprocessed training data into training and validation sets, initialised the neural network architecture and the hyperparameters, and trained the model on the training set with back-propagation and gradient descent. Then, they used grid or random search to maximise the model's performance and then verified the trained model with cross-validation to test its generalisation performance and robustness.

- Evaluation and Interpretation
  The performance metrics used in the research paper to evaluate the performance of the CreditNetXAI were accuracy, which measures the proportions of correct predictions among all predictions made by a classifier; sensitivity, which measures the proportion of actual positive instances correctly identified; and specificity, which measures the proportion of actual negative instances correctly identified.

2.3.3 Critical Analysis

The CreditNetXAI model produced promising results, with an accuracy of 0.8350, suggesting that it properly classified the majority of credit applications. Furthermore, the model's sensitivity of 0.8823 indicated that it could detect most of the true positive cases. The model accurately detected most of the true negative situations: it is visible by looking at the specificity value of 0.9879.

These findings indicate that the CreditNetXAI model can be a useful tool for credit risk assessment. It is highly accurate and can detect both positive and negative instances with high sensitivity and specificity. As it is visible from the below table [5], in all performance criteria, the model outperformed alternative linear and nonlinear models typically used in credit scoring, such as XGBoost and support vector machine.



Model	Accuracy	Sensitivity	Specificity
CreditNetXAI	0.8350	0.8823	0.9879
XGB	0.7978	0.8429	0.7057
SVM	0.7916	0.8410	0.6906
RF	0.7928	0.8521	0.6717
NN	0.7531	0.8281	0.6000

Figure 1: Table regarding performance metrics

The research paper explores all the different phases behind the realisation of the model in a comprehensible format making it accessible even for readers that are just approaching the topic of the neural networks.

In the paper, the authors explain how challenges may arise when applying the CreditNetXAI model to different datasets or real-world scenarios due to data variability, cultural and regional financial differences, and evolving financial patterns.

For the future, they propose to apply the CreditNetXAI's algorithm to address various challenges in the world of finance, but also other sectors as well.

2.4 Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using Lending Club data

2.4.1 Summary

Credit scoring is the process of placing someone on a creditworthiness scale. This scale is significant because it guides judgments about granting credit to applicants, which is especially useful in peer-to-peer transactions. Peer-to-peer (P2P) financing is classified as commercial crowdfunding, in which internet platforms collect cash from the "crowd" to collectively finance higher-value loans for individuals or enterprises. The main distinction between bank loans and peer-to-peer loans is that

[5] The table has been taken from the article [Talaat et al., 2024]

the latter are unsecured, which means there is no collateral behind them: if the borrower is unable to repay the amount, the lender institution will not receive anything back. This explains why P2P lending is risky for investors.

To address this, the researchers attempted to construct an artificial neural network framework to assess default prediction in P2P loans, which was then compared to other models to ensure its reliability [Chang et al., 2022].

2.4.2 Methodology

The research data was sourced from the Lending Club's accessible P2P lending data. The major features analyzed include the borrower's personal information, loan purpose, credit history, debt status, and others, for a total of 16 features, with the most essential ones pre-trained using the XGBoost approach. Because the loans after 2015 had not fully expired when the data were obtained in 2018, it was impossible to determine whether they were in default. As a result, the researchers used three-year loan data throughout a two-year period (January 2013 to December 2014), totaling 282763 loans.

The variable "`loan_status`" served as the default label, with "fully Paid" denoting no default and "Late", "Default", and other values indicating defaults. There were 245332 non-default loans (86.8%) and 37,431 default loans (13.2%).

They trained seven different models in total (logistic regression, support vector machine, decision tree, random forest, XGBoost, LightGBM, and Neural Network), and then compared the performance of the artificial neural network to these other models by evaluating performance measures such as accuracy, recall, precision, and F1.

2.4.3 Critical Analysis

In terms of accuracy and F1 score, the two most thorough indicators, XGBoost outperformed LightGBM and random forest, both of which had an accuracy of around 87.6%. The logistic regression and 2-layer neural network algorithms had the worst results. This demonstrates that these two models are less suitable for credit scoring in terms of predictive power. In terms of precision, logistic regression tested the best, with a precision of more than 86.5%, but XGBoost and LightGBM also had high precision values, above 85%.

In general, comparing the predictive performance of each model revealed that gradient-boosting decision tree methods, such as XGBoost and LightGBM, are the best P2P credit-scoring models, outperforming 2-layer neural networks and traditional logistic regression.

The study demonstrates how new models can improve the performance of traditional models such as logistic regression; however, among the new models, in the case of P2P transactions, the neural network model proposed by these researchers does still not perform well enough to be implemented.

Similarly to a previous article examined in this paper, there is no explanation of how they constructed each model, only a theoretical description of what each model does. The authors suggest a possible improvement: because the data contains description features, such as the reasons for the loans and the lender's credit document, the textual information can be converted into a numerical form using natural language processing, like sentiment analysis, to obtain more information for default classification.

2.5 Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models

2.5.1 Summary

To reduce greenhouse gas emissions, a long-term shift to low-carbon energy sources is required. Although awareness of the need for a circular economy has increased significantly, much more work needs to be done to accomplish sustainability goals.

The Capex investment has prompted a low-carbon corporate shift and is anticipated to put financial flexibility under pressure. Financial flexibility is determined by credit ratings, with higher credit ratings resulting in reduced borrowing costs. As a result, the availability of external credit ratings can effectively reduce carbon emissions by allowing the necessary capital expenditure. Forecasting such ratings will help to develop a financial strategy.

In this work, the researchers assessed the performance of a neural network model and other three machine learning models to predict the credit ratings of environmentally friendly enterprises in the Eurozone. There were two reasons for choosing the Eurozone. First, the European Union has been in the forefront of the shift to sustainable business models to cut greenhouse gas emissions. Second, currency uniformity is essential for comparing credit capacity between countries [Yu et al., 2022].

2.5.2 Methodology

The Carbon Disclosure Project (CDP) provides a grading system for ranking businesses' progress toward climate change mitigation and low-carbon corporate strategies. The rankings, which are based on annual disclosures and pertain to environmental management, are classified into five categories from A to E. The CDP ranks 13 of the 19 member states based on their environmental performance. As a result, the sample included firms from these 13 countries with A or B credit ratings assigned by a respected rating agency and made publicly available. Based on this, the final sample included 335 non-financial firms, with a ten-year sample duration chosen to provide enough data for learning and forecasting.

They divided it into two phases: learning and prediction (2017–2019). Additionally, quarterly data for macro and micro factors were obtained from Thomson Reuters Eikon and Bankscope. Furthermore, credit ratings were obtained from the appropriate credit rating organizations' websites and divided into three categories: investment grade, speculative grade, and default. These were categorized as CR1 (Maximum) through CR7 (Default), with numerical values ranging from 17 (AAA) to 1 (Default).

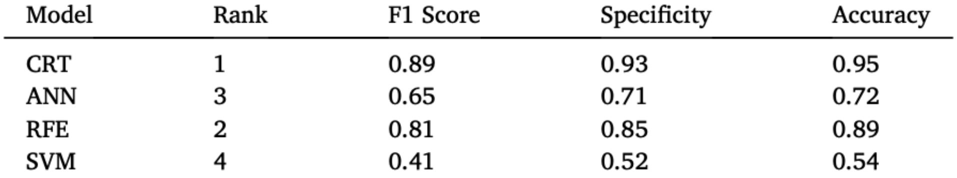
Credit rating data, as well as financial and macroeconomic inputs, were utilized to train machine learning models, and the F1-score, specificity, and accuracy were used to assess forecast accuracy.

Four models were used: classification and regression trees (CRT), artificial neural networks (ANN), random forest ensemble (RFE), and support vector machines (SVM).

To train the artificial neural network model, they began by randomly assigning weights to all of the inputs in the first node. These were used to estimate the output of each node and, eventually, the target. The within-sample error was determined by comparing the objective and actual values. Finally, the error was returned to each node, and the relevant contribution was calibrated and adjusted. They used three hidden layers and a decay factor of 0.1 to carry out the assessment.

2.5.3 Critical Analysis

The four machine learning models produced similar findings for the macro and micro variable rankings. A company's capital structure is highly valued, however, financial risk is highly valued for enterprises with a high likelihood of default. Although the variable rankings are comparable across models, the situation differs when it comes to rating prediction. As, we can see from the below table [6], the metrics showed that CRT was the best of the four models, with the highest F1 score, specificity, and accuracy. The RFE is the second-best model, after ANN and SVM.



Model	Rank	F1 Score	Specificity	Accuracy
CRT	1	0.89	0.93	0.95
ANN	3	0.65	0.71	0.72
RFE	2	0.81	0.85	0.89
SVM	4	0.41	0.52	0.54

Figure 2: Table regarding performance metrics

CRT's ability to forecast the ratings of organisations with fewer emissions is due to the fact that the data does not need to be linear or parametric. Furthermore, since the output (credit ratings) is a class-based continuous variable, CRT appears to be a good fit. These findings are novel in terms of carbon-neutral enterprises because they demonstrate how credit rating forecasts may be done using models other than random forest algorithms, which have been used in previous studies, and without utilising a complicated model such as an artificial neural network. The research argues that the neural network model is not the best answer for assessing the risk of default of a decarbonized firm.

This is intriguing research since it helps us to reflect on the true necessity of employing neural network models. The previous analysed papers used the scenario of a bank assessing the risk of issuing money or a credit card to a consumer, whereas the writers of this paper intended to evaluate the creditworthiness linked with a corporation. The findings of the investigation from both perspectives demonstrated that neural networks do not yet provide a level of performance sufficient enough for regular implementation by banks or credit rating institutions.

The research was well-written and easy to grasp. I appreciated how the authors described a possible strategy to improve the model by taking into account both quantitative and qualitative information.

[6] The table has been taken from the article [Yu et al., 2022]

3 Strengths and Limitations

Article	**Strengths**	**Limitations**
Deep Neural Networks for Behavioral Credit Rating	The authors were able to demonstrate how the neural network is able to obtain better results than those obtained using linear models	The paper does not describe the process that was done to train all the different models. In addition, there is no sign of possible limitations and future path that could be pursued
Secured Loan Prediction System using Artificial Neural Network	The paper details with great clarity all of the processes that were taken to create the neural network and it demonstrates how the model has a lot of opportunity for development even though it still performed better than the linear models it was compared to	The big limitation of the study is that the criterion used to evaluate the model was just the accuracy
Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction	The research paper explores all the different phases behind the realisation of the model in a comprehensible format. It also describes the limitations encountered in the study and the future work that can be done to improve the model	I did not find particular limitations in the research paper
A Hybrid Approach for Predicting Probability of Default in Peer-to-Peer (P2P) Lending Platforms Using Mixture-of-Experts Neural Network	The paper is well written and easy to understand. I appreciated the inclusion of some hypotheses on how to improve the model	There is no explanation of how they constructed each model. In addition, the researchers do not talk about the limitations of their work and the difficulties they have encountered
Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models	The research was well written and easy to understand. I appreciated how the researchers described a possible strategy to improve the model	The main limitation has been that they described deeply only the procedure to create the neural network model

Table 1: Summary of Reviewed Articles

4 Conclusion

Financial institutions and investors heavily rely on models that are able to assess the creditworthiness of their clients. The increase in available data has pushed these institutions to develop new and more advanced models able to understand the complex relationship between variables.

Among these new models, a particular focus has been dedicated to the neural network, and the objective of this paper has been to analyze five different pieces of research that could explain whether this model is the optimal solution. This has been achieved by comparing the performance metrics of the neural network to the results obtained through other linear and non-linear models.

As highlighted by the results achieved in the five papers analyzed, the neural network can perform at a higher level than the linear models; however, it never represents the best model, with the only exception of the second paper [Adebiyi et al., 2022], where, however, the results are slightly biased since they considered only one performance metric. For example, in the first paper [Merćep et al., 2021], the researchers demonstrated how the deep neural network model for behavioral credit risk assessment can obtain better results compared to the linear benchmark model, but the results obtained by using a gradient boosting-based model were better. Similarly, the fourth article [Chang et al., 2022]demonstrates how new models can improve the performance of traditional models such as logistic regression; however, among the new models, in the case of P2P transactions, the neural network model proposed is not able to perform well enough to be implemented when compared to models such as XGBoost and LightGBM. Finally, in the fifth paper [Yu et al., 2022], the results show how the classification and regression trees and random forests perform better.

However, all the authors suggest further research on the subject since the improvement of the model could be beneficial to both banks and consumers. A possible way of improving the performance of the neural network and increasing its interpretability, as shown in the third article [Talaat et al., 2024], is through the use of explainable artificial intelligence (XAI) techniques that could help build trust in the models and facilitate their adoption in real-world applications.

Financial institutions around the world have understood the need to implement new and more sophisticated models to get high-quality results, and they intend to invest a large quantity of money to achieve this objective. However, as of right now, there is still a long way to go before being able to implement the neural network models in credit scoring.

References

[Adebiyi et al., 2022] Adebiyi, M. O., Adeoye, O. O., OGUNDOKUN, R. O., Okesola, J. O., and Adebiyi, A. A. (2022). Secured loan prediction system using artificial neural network. *Journal of Engineering Science and Technology*, 17(2):0854–0873.

[Chang et al., 2022] Chang, A.-H., Yang, L.-K., Tsaih, R.-H., and Lin, S.-K. (2022). Machine learning and artificial neural networks to construct p2p lending credit-scoring model: A case using lending club data. *Quantitative Finance and Economics*, 6(2):303–325.

[Merćep et al., 2021] Merćep, A., Mrčela, L., Birov, M., and Kostanjčar, Z. (2021). Deep neural networks for behavioral credit rating. *Entropy*, 23(1).

[Talaat et al., 2024] Talaat, F. M., Aljadani, A., Badawy, M., and Elhosseini, M. (2024). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications*, 36(9):4847–4865.

[Yu et al., 2022] Yu, B., Li, C., Mirza, N., and Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, 174:121255.