**CHAPTER TWO**

**LITERATURE REVIEW**

## **Introduction**

Any research builds on the existing knowledge in a particular field. This chapter reviews various prior literatures on Credit Risk Management of Commercial Banks in Tanzania using Predictive modeling Techniques. Credit risk management plays a pivotal role in the banking sector, as financial institutions face the challenge of assessing the creditworthiness of borrowers to make informed lending decisions.Traditionally, credit risk assessment has relied on manual processes and subjective judgment, which are often time-consuming and prone to human biases. Techniques such as weight‐of‐evidence measure, regression analysis, discriminant analysis, probit analysis, logistic regression, linear programming, Cox’s proportional hazard model, support vector machines, decision trees, neural networks, K‐nearest neighbor (K‐NN), genetic algorithms, and genetic programming are some of statistical credit scoring methods used by researchers, credit analysts, and lenders until recent (Abdou & Pointon, 2011).However, recent advancements in machine learning (ML) techniques have opened up new avenues for improving credit risk management by enabling the development of predictive modeling approaches. This literature review explores the existing research and studies that focus on enhancing credit risk management in banking using machine learning. Finally, relevant and related empirical works will be reviewed and the knowledge gap will be identified.

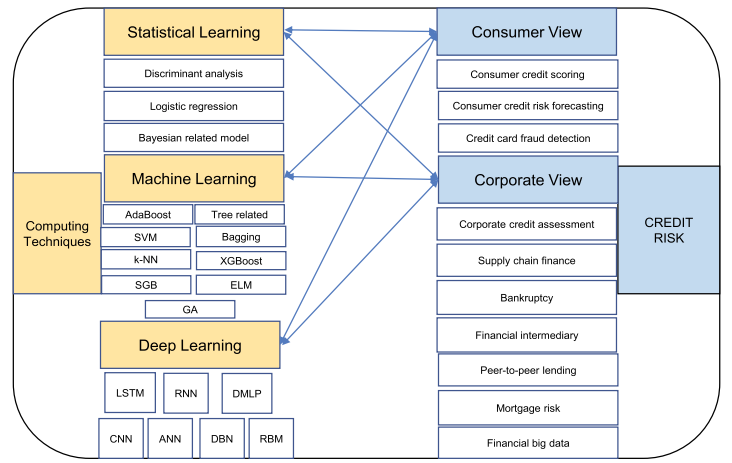


Figure 1. Taxonomy of approaches and algorithms (Shi et al., 2022)

**2.2 Credit scoring**

Credit scoring is a crucial aspect of credit risk management in the banking industry. It involves classifying clients into different risk categories, typically categorized as "good" and "bad" borrowers. Traditionally, credit scoring has been approached as a classification problem, aiming to accurately predict the creditworthiness of borrowers based on various factors. One technique that has been applied to credit scoring is the use of genetic algorithms. Genetic algorithms utilize a fitness function to evaluate the predictive performance of different solutions or individuals in

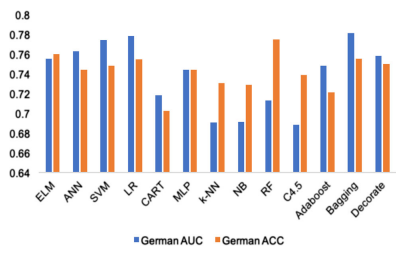
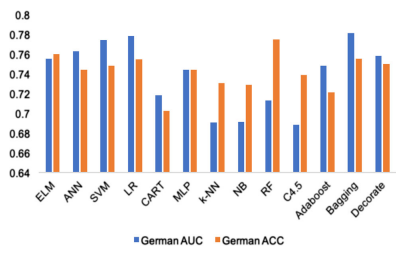
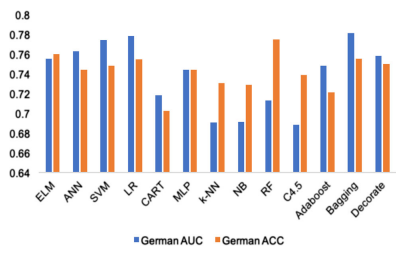
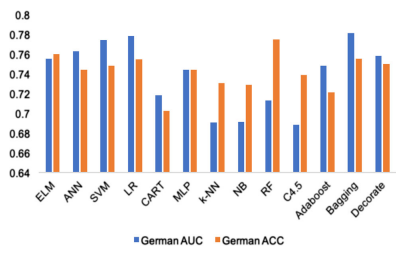
the population. In the context of credit scoring, each individual represents a genetically encoded solution to the classification problem, and its fitness score reflects its ability to provide an accurate credit risk assessment. The genetic algorithm operates by evolving an initial population of solutions into a new population. This evolutionary process involves genetic-inspired operations, such as mutation and crossover, to generate new individuals with potentially improved fitness scores. The goal is to find the most optimal combination of features or variables that can effectively classify borrowers into the appropriate risk categories (Vanve & Patil, n.d.). Furthermore, credit scoring can also be approached from a profit-based perspective. In this approach, the classification of customers is associated with potential profit or loss. Correctly classifying borrowers as good or bad can lead to profit, while incorrect classification may result in potential profit loss. This profit-based approach considers the financial implications of credit risk assessment, emphasizing the importance of accurately categorizing borrowers to maximize profitability.

**2.3 The Evolution of Credit Risk Management:**

Credit risk management has undergone significant transformations over the years, from traditional rule-based approaches to more sophisticated models that leverage statistical and ML techniques. This section provides an overview of the evolution of credit risk management practices, highlighting the challenges faced and the need for improved predictive modeling techniques.

**2.4 Machine Learning in Credit Risk Management:**

This section delves into the applications of machine learning in credit risk management. It examines the various ML algorithms and methodologies that have been employed to predict credit risk, including logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks (Shi et al., 2022).



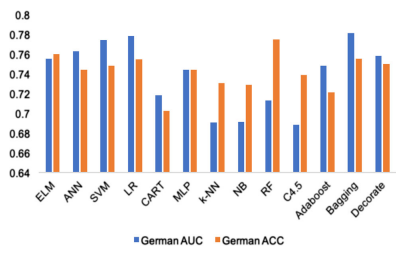


Figure 2. The accuracy for Germany credit data

**2.5 Utilization of Alternative Data Sources**

To enhance credit risk management, researchers and practitioners have explored the use of alternative data sources beyond traditional financial information. This section reviews studies that have incorporated non-traditional data, such as social media activity, transactional data, and mobile application data, into credit risk modeling. The potential benefits and challenges associated with leveraging alternative data sources are examined, along with the impact of such data on credit risk prediction accuracy (Simão, n.d.).

**2.6 Integration of Economic Effects**

Credit risk is not solely influenced by individual borrower characteristics but is also affected by broader economic factors. This section explores research that focuses on integrating economic effects, such as unemployment rates, inflation, and housing market indicators, into credit risk modeling. The findings highlight the importance of considering macroeconomic factors in credit risk assessment and demonstrate how their inclusion can improve the accuracy and robustness of predictive models (Simão, n.d.).

**2.7 Model Explainability and Interpretability**

The interpretability of credit risk models is crucial for regulatory compliance and transparency. This section examines research that addresses the challenge of model explainability in machine learning-based credit risk management. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and fuzzy logic are explored for their potential to provide transparent and understandable explanations for credit risk predictions.

## **2.8 Related work**

The paper (Simão, n.d.) has been a closely related work that builds up a major contribution for the research I am conducting in the following aspect:-

Historical Development: The concept of credit scoring dates back to the 1940s, and Linear Discriminant Analysis was one of the early techniques used for assessing credit risk. Since then, various developments have been proposed in the literature, including Logistic Regression, Survival Analysis, and machine learning methods.

Machine Learning in Credit Scoring: Many studies have explored the application of machine learning techniques in predicting loan default and credit risk. These techniques include Support Vector Machines (SVM), Genetic Algorithms, Neural Networks, Random Forests, Decision Trees, AdaBoost, and ensemble methods. The performance of these models has been evaluated using metrics such as accuracy, AUC, Brier Score, and Type I and Type II errors.

Comparative Studies: Several comparative studies have been conducted to assess the performance of different models. These studies have shown that ensemble methods, such as stacking, bagging, and boosting, often outperform individual classifiers. Tree-based models, such as Random Forest and XGBoost, have demonstrated good performance and stability. It has been found that machine learning models, such as boosted regression trees, can provide better predictive performance for mortgage credit risk compared to logistic regression.

Important Variables: Common variables found to be important across multiple studies include income, loan amount, loan purpose, employment status, and home ownership. These variables have been consistently used to predict loan default and assess creditworthiness.

Performance Evaluation: Various performance evaluation criteria have been used to compare different models, such as accuracy, AUC, Brier Score, H-measure, and Kolmogorov-Smirnov statistics. Heterogeneous ensemble classifiers have been found to perform well, outperforming logistic regression in many cases.

**2.9 Research gap**

Despite the advancements made in credit scoring modeling, there exist several limitations and areas for future research. One significant research gap is the exploration of alternative data sources and the integration of economic effects to improve the accuracy and interpretability of credit scoring models (Teng & Lee, 2019).

Currently, the availability of sensitive and confidential data from financial institutions poses a challenge, leading to limited access for researchers (Simão, n.d.). Therefore, this research focuses on identifying and utilizing new data sources, such as the vast amount of digital information recorded on social networks and mobile applications. This data, when analyzed from a behavioral perspective, may provide valuable insights for consumer credit risk research.

Additionally, while personal characteristics play a crucial role in determining borrowers' creditworthiness, the external economic environment also impacts loan performance. Factors such as unemployment rates and house prices can influence borrowers' ability to repay loans. Hence, this research investigates the integration of detailed economic information into credit scoring models, aiming to improve their accuracy and interpretability by considering the broader economic context.

Another aspect that requires attention is the model explainability. The decisions made based on credit scoring models should be transparent and operate within equal opportunity laws. Although variable-importance scores provide some insight into the predictors' significance, a deeper exploration of model explainability is necessary. Techniques like LIME (Local Interpretable Model-Agnostic Explanations) and fuzzy logic have shown potential in enhancing the explanatory capability of predictive models. Exploring these methods, along with other interpretability frameworks, can contribute to developing credit scoring models that are more transparent and explainable.