IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

EXPLAINABLE AI (XAI) FOR LOAN APPROVAL AND CREDIT RISK PREDICTION

Dr. Mohammed Sharfuddin, [1] Professor, Department of Computer Science and Technology,

Deccan College of Engineering and Technology, Hyderabad, India

Nadia Mujeeb,[2] PG Scholar, Department of Computer Science and Engineering,

Deccan College of Engineering and Technology, Hyderabad, India

ABSTRACT – Financial institutions across the world rely heavily on machine learning models to decide whether a loan application should be approved or rejected. Models such as XGBoost (eXtreme Gradient Boosting) give very accurate predictions on this kind of structured, tabular data, which is exactly why banks and non-banking financial companies have started using them at a large scale for credit scoring. The problem is that these models work like a black box. They give a decision but they do not explain why that decision was made, which leaves both the loan applicant and the bank employee with no clear reason behind an approval or a rejection. This project tries to solve that exact problem by building a complete Explainable Artificial Intelligence, or XAI, based credit risk prediction system.

The system combines a well-tuned XGBoost classifier with three different explanation techniques, namely SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations) and DiCE (Diverse Counterfactual Explanations) counterfactuals, so that every prediction made by the model can be explained from more than one angle. The model is trained and tested on the German Credit dataset, which has been extended to ten thousand records so that the training process is more robust and closer to a real world scenario. Before training, the raw data goes through a proper preprocessing pipeline that handles categorical encoding, feature scaling and class balancing using SMOTE, since the number of defaulters in the original dataset is quite low compared to the number of approved applicants.

Once the model is trained, SHAP is used to find out which features matter the most across the whole dataset as well as for one particular applicant, LIME is used to build a small local model around a single prediction so that it becomes easy to interpret, and DiCE is used to tell a rejected applicant the smallest possible change in their profile that would flip the decision to an approval. All three of these explanation methods are also benchmarked against each other on faithfulness, stability and the time they take to compute, so that the strengths and weaknesses of each method are clearly visible rather than just assumed.

Apart from explainability, this project also checks whether the model is being fair to different age groups. Fairlearn and AIF360 are used to compute metrics such as demographic parity difference, equalised odds difference and disparate impact ratio, and a reweighing step is applied afterwards to reduce any bias that is found. Finally, the entire pipeline, starting from data loading and going all the way to explanation generation and fairness auditing, is packaged inside an interactive, multi tab Streamlit dashboard where a user can upload an applicant's data and instantly see the prediction, the explanation behind it and the fairness status of the model. The final system supports the right to explanation principle mentioned in data protection laws like the GDPR and shows that high accuracy and model transparency are not two opposite goals, they can be achieved together using tools that are completely open source and free to use.

Index Terms —Credit Risk Prediction, Explainable Artificial Intelligence (XAI), Machine Learning, XGBoost, SHAP, LIME, DiCE, Counterfactual Explanations, Fairness-Aware Machine Learning, Fairness Evaluation, Bias Mitigation, German Credit Dataset, Reweighing, Predictive Analytics, Streamlit.

1. INTRODUCTION

1.1 Background

Credit risk assessment is a fundamental process in the financial sector because it assists financial institutions in evaluating the potential risk associated with lending to applicants. Traditional credit assessment methods generally rely on historical financial records, statistical techniques, predefined rules, and manually selected applicant characteristics. With the rapid development of **Artificial Intelligence (AI)** and **Machine Learning (ML)**, automated prediction systems are increasingly capable of analyzing large numbers of financial and demographic features to identify patterns associated with creditworthiness and potential repayment risk. These techniques can improve the efficiency and consistency of credit-risk assessment; however, the use of complex machine-learning models also introduces important concerns related to transparency, interpretability, and fairness.

Many advanced machine-learning algorithms are capable of achieving strong predictive performance on structured financial data. However, models with complex internal decision processes may provide limited information about why a particular prediction has been generated. In credit-risk applications, this limitation is significant because automated predictions can influence decisions that affect individuals. When an applicant is classified into a particular credit-risk category, users, analysts, and other stakeholders may need to understand the factors that contributed to the prediction. A prediction without an understandable explanation can reduce trust in the model and make it difficult to examine whether the system is relying on reasonable patterns.

Explainable Artificial Intelligence (XAI) addresses this challenge by providing methods for interpreting machine-learning models and individual predictions. XAI techniques can identify the features that contribute to a model's output and provide additional information about how a decision has been reached. In this study, three complementary explanation approaches are considered: SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and DiCE (Diverse Counterfactual Explanations). SHAP provides feature-attribution information and can be used to examine both global model behavior and individual predictions. LIME generates local explanations by approximating the behavior of a complex model around a specific instance. DiCE provides counterfactual explanations by identifying alternative feature conditions that may result in a different model prediction. Together, these approaches provide multiple perspectives for understanding automated credit-risk decisions.

In addition to explainability, **fairness-aware machine learning** is an important consideration in financial decision-support systems. Machine-learning models learn from historical data, and such data may contain imbalances or patterns that result in differences between demographic groups. Consequently, a model can demonstrate good overall predictive performance while still producing unequal outcomes for particular groups. Therefore, conventional performance measures such as accuracy, precision, recall, F1-score, and ROC-AUC alone may not provide a complete evaluation of a credit-risk prediction system.

The proposed research presents an Explainable and Fair XGBoost-Based Credit Risk Prediction Framework Using SHAP, LIME and DiCE. The system uses **XGBoost** as the primary machine-learning model because of its effectiveness in handling structured and tabular datasets. The proposed framework integrates data preprocessing, feature preparation, class balancing, model training, threshold optimization, explainable AI, counterfactual reasoning, fairness evaluation, and bias mitigation within a unified workflow.

The study uses the **German Credit dataset** as the basis for model development and experimentation. The dataset contains financial and personal attributes associated with credit-risk classification. Before model training, the data undergoes preprocessing operations such as cleaning, categorical feature transformation, feature preparation, and class balancing. The prepared data is subsequently used to train the XGBoost classification model. Model performance is evaluated using commonly used classification measures to determine its ability to distinguish between the defined credit-risk categories.

The proposed framework further incorporates a quantitative comparison of the explanation techniques. Rather than assuming that all explanations are equally useful, SHAP, LIME, and DiCE are evaluated according to criteria such as

faithfulness, stability, computational time, and actionability. This comparison is important because an explanation method that is highly consistent may require additional computation, while a faster explanation may provide less stable or less faithful results. Similarly, counterfactual explanations can provide actionable information but may involve higher computational complexity.

The complete system is integrated into an interactive **Streamlit-based application** that provides functionalities for model prediction, explanation, counterfactual generation, fairness analysis, and comparative visualization. Users can examine individual predictions and explore the factors associated with model decisions using different XAI techniques. The interface is intended to demonstrate how predictive modeling, interpretability, and fairness analysis can be presented together in a practical decision-support environment.

Overall, the proposed study integrates Machine Learning, XGBoost, Explainable Artificial Intelligence, SHAP, LIME, DiCE, Counterfactual Explanations, Fairness Evaluation, Bias Mitigation, and Interactive Visualization into a unified credit-risk prediction framework. Rather than focusing exclusively on predictive performance, the research evaluates the proposed system from the perspectives of prediction quality, explanation capability, computational efficiency, and fairness. The objective is to contribute to the development of more transparent and systematically evaluated machine-learning-based credit-risk systems.

1.2 Research Gap

Existing research has demonstrated the effectiveness of machine-learning techniques for credit-risk prediction, and recent studies have increasingly investigated the use of explainable AI to improve the transparency of automated decisions. However, many existing approaches primarily focus on achieving high predictive performance or employ only a single explanation technique. While individual methods such as SHAP or LIME can provide useful insights, they may differ in their ability to accurately represent model behavior, maintain consistent explanations, generate results efficiently, or provide actionable information to users.

A further gap exists between model explainability and fairness-aware evaluation**.** An explanation can indicate which features influenced a particular prediction, but it does not necessarily determine whether the model behaves consistently across different demographic groups. Similarly, fairness metrics can identify potential disparities between groups but may not provide detailed information about the factors contributing to individual predictions. Therefore, explainability and fairness should be considered as complementary components of a responsible machine-learning framework.

Therefore, the proposed research addresses the need for an integrated explainable and fairness-aware credit-risk prediction framework. The study combines XGBoost-based prediction with SHAP, LIME, and DiCE explanations, evaluates the quality of these methods using multiple criteria, examines potential demographic disparities through fairness metrics, and applies reweighing as a bias-mitigation strategy. The integration of these components within an interactive Streamlit-based system distinguishes the proposed framework from approaches that focus only on prediction accuracy or a single dimension of model evaluation.

1.3 Research Objectives

The study focuses on the following objectives:

1. **To develop an XGBoost-based credit-risk prediction model** capable of classifying applicants into the defined credit-risk categories.
2. **To preprocess and prepare the credit-risk dataset** using appropriate data cleaning, feature transformation, encoding, and class-balancing techniques.
3. **To integrate SHAP, LIME, and DiCE** for providing global, local, and counterfactual explanations of model predictions.
4. **To evaluate and compare the XAI techniques** based on faithfulness, stability, computational time, and actionability.
5. **To evaluate fairness in model predictions** using measures such as Disparate Impact, Demographic Parity Difference, Equalized Odds Difference, and Accuracy Difference.
6. **To apply a reweighing-based fairness mitigation technique** and examine its influence on prediction disparities.

7. **To develop an interactive Streamlit-based application** for credit-risk prediction, explanation generation, counterfactual analysis, and fairness visualization.

8. **To evaluate the proposed framework** based on predictive performance, interpretability, fairness, computational efficiency, and practical usability.

1.4 Limitations of the Study

The proposed framework has several limitations. First, the study is based on the **German Credit dataset,** and the experimental dataset is extended through augmentation. Therefore, the resulting experimental performance may not directly represent the behavior of the system on large-scale, contemporary, real-world financial data.

Second, the proposed system focuses primarily on **age as the selected protected attribute** for fairness evaluation. Other demographic or socio-economic attributes may also influence model behavior but are not extensively analyzed within the current implementation. In addition, fairness is evaluated using a selected group of metrics, and different fairness definitions may produce different interpretations of model behavior.

Finally, the proposed system is designed as an **experimental decision-support framework** and is not intended to function as a fully autonomous credit-approval system. Real-world deployment would require additional validation, regulatory compliance, data-security mechanisms, continuous monitoring, and expert human oversight.

1.5 Rationale of the Study

The increasing use of machine-learning models in financial decision-making creates a need for systems that are not only predictive but also understandable and systematically evaluated for fairness. Although high-performance algorithms can improve the efficiency of credit-risk assessment, predictions produced by complex models may be difficult for users and stakeholders to interpret. In addition, strong predictive performance alone does not guarantee that a model produces equitable outcomes across different groups.

The proposed study addresses these challenges by combining **XGBoost-based prediction,** Explainable Artificial Intelligence, counterfactual reasoning, fairness evaluation, and bias mitigation within a unified framework. The use of SHAP, LIME, and DiCE enables the system to provide complementary forms of explanation, while quantitative comparison helps examine their practical strengths and limitations. Fairness analysis and reweighing further extend the evaluation beyond conventional classification metrics.

Overall, the study is motivated by the need for a more transparent, explainable, and fairness-aware approach to credit-risk prediction. By integrating prediction, explanation, fairness assessment, mitigation, and interactive visualization, the proposed framework aims to demonstrate a broader approach to evaluating machine-learning systems used in high-impact decision-support environments.

2. LITERATURE REVIEW

The increasing use of machine-learning techniques in financial decision-making has led to growing research interest in credit-risk prediction, model interpretability, and fairness. Conventional credit-risk models can provide useful predictions, but complex machine-learning approaches may make it difficult for users to understand the reasoning behind individual outcomes. Explainable Artificial Intelligence (XAI) has therefore become an important research area for improving transparency, while fairness-aware machine learning has been investigated to identify and reduce undesirable differences between demographic groups.

2.1 Machine Learning for Credit Risk Prediction

Machine-learning methods have been widely applied to credit-risk prediction because they can identify nonlinear relationships among multiple applicant characteristics. Tree-based ensemble methods are particularly suitable for structured financial datasets because they can model complex feature interactions and handle different types of input variables.

Among these techniques, **XGBoost** is an effective gradient-boosting algorithm that combines multiple decision trees to produce a strong predictive model. Its regularization mechanisms and ability to model nonlinear relationships make it suitable for classification tasks involving structured credit data. However, the predictive capability of XGBoost comes with an interpretability challenge because the final prediction is generated through the combined behavior of multiple trees.

Therefore, although XGBoost can provide strong predictive performance, additional explanation techniques are required when the objective is to understand the factors influencing individual predictions. This motivates the integration of XAI methods with the predictive model.

2.2 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) refers to methods that provide understandable information about the behavior and predictions of machine-learning models. Adadi and Berrada highlighted the growing importance of explainability as machine-learning systems are increasingly used in decision-making applications. XAI can improve transparency by identifying important features, explaining individual predictions, and helping users understand model behavior.

In credit-risk prediction, explainability is particularly relevant because users may need to understand why a record has been classified into a particular risk category. Explanations can also assist developers and analysts in identifying unexpected model behavior and evaluating whether the model is

2.3 SHAP-Based Explanations

SHAP (SHapley Additive exPlanations) is a feature-attribution approach based on Shapley values from cooperative game theory. It estimates the contribution of individual features to a model prediction relative to a reference output. SHAP can be used to generate both global explanations, which describe overall feature importance, and local explanations, which describe the factors influencing a particular prediction.

In credit-risk applications, SHAP can help identify which applicant characteristics contribute positively or negatively to a predicted credit-risk category. This provides a more detailed interpretation than simply presenting the model output. The proposed framework uses SHAP to examine feature contributions and to support interpretation of individual and overall model behavior.

2.4 LIME-Based Explanations

LIME (Local Interpretable Model-agnostic Explanations) explains an individual prediction by creating perturbed samples around the selected instance and fitting an interpretable local model. The resulting approximation provides an indication of which features have the greatest influence on that particular prediction.

LIME is useful when a user needs a localized explanation for a specific record. However, because it relies on local approximation and sampling, explanations can vary depending on the generated neighborhood and sampling process. Consequently, the proposed study evaluates LIME not only for its explanatory capability but also in terms of stability and computational efficiency.

2.5 Counterfactual Explanations Using DiCE

DiCE (Diverse Counterfactual Explanations) provides counterfactual explanations by generating alternative versions of an input record that may lead to a different model prediction. Unlike feature-attribution methods that primarily explain why a prediction occurred, counterfactual explanations focus on what could potentially change the prediction.

Counterfactual reasoning can provide more actionable information because it presents alternative feature conditions associated with a different model outcome. However, generating multiple diverse counterfactual examples can require additional computation. The proposed framework therefore incorporates DiCE alongside SHAP and LIME to provide a complementary form of explanation and evaluates its actionability and computational requirements.

2.6 Fairness in Machine Learning

While explainability focuses on understanding model decisions, **fairness-aware machine learning** focuses on examining whether automated systems produce undesirable disparities between groups. Mehrabi et al. discussed different sources and definitions of bias in machine-learning systems and highlighted the importance of evaluating fairness in automated decision-making.

In credit-risk prediction, fairness is particularly important because historical data may contain demographic differences that can influence model outcomes. A model may achieve satisfactory overall predictive performance while still producing different outcomes for different groups. Therefore, fairness evaluation should complement conventional performance metrics.

The proposed framework evaluates fairness using Disparate Impact, Demographic Parity Difference, Equalized Odds Difference, and Accuracy Difference. These measures provide different perspectives on potential disparities in model outcomes. The study also applies a **reweighing-based mitigation technique** to investigate whether fairness-related differences can be reduced.

2.7 Integrated Explainability and Fairness

Existing research has explored credit-risk prediction, explainability, and fairness as individual areas, but integrating these components into a single practical framework remains challenging. An XAI method can explain the contribution of features without determining whether outcomes are fair, while fairness metrics can identify group-level disparities without explaining individual predictions.

The proposed work addresses this issue by integrating XGBoost prediction, SHAP, LIME, DiCE, quantitative XAI evaluation, fairness assessment, and fairness mitigation within one framework. The system additionally provides a Streamlit-based interface for interacting with the prediction and analysis components.

2.8 Research Gap

The reviewed literature indicates several limitations in existing machine-learning-based credit-risk prediction systems. Many approaches focus mainly on **predictive performance**, using measures such as accuracy, precision, recall, F1-score, and ROC-AUC, while providing limited information about how individual predictions are produced.

Another gap concerns the integration of **explainability and fairness evaluation** within a single credit-risk prediction framework. While methods such as SHAP and LIME can identify influential features, they do not directly determine whether model outcomes are equitable across demographic groups. Similarly, fairness metrics can identify differences between groups but do not explain the factors responsible for individual predictions. Counterfactual methods such as DiCE provide an additional perspective by showing alternative input conditions that may influence a model outcome, but they are not always integrated with feature-attribution methods and quantitative fairness analysis. Therefore, there is a need for a unified framework that combines XGBoost-based prediction, SHAP, LIME, DiCE, quantitative XAI evaluation, fairness assessment, bias mitigation, and interactive visualization**.**

3. RESEARCH METHODOLOGY

The proposed Explainable and Fair XGBoost-Based Credit Risk Prediction Framework Using SHAP, LIME and DiCE follows a systematic approach for developing and evaluating a machine-learning-based credit-risk prediction system. The methodology integrates data preprocessing, class balancing, XGBoost classification, threshold optimization, explainable artificial intelligence (XAI), fairness evaluation, bias mitigation, and interactive visualization within a unified framework. The system uses the German Credit dataset as the foundation for experimentation and classifies records into **Good Credit Risk** and **Bad Credit Risk** categories. The framework additionally generates explanations for individual predictions and evaluates potential disparities across demographic groups.

The overall workflow is:

German Credit Dataset → Data Cleaning → Feature Preparation → Encoding → Class Balancing → Train/Test Split → XGBoost Model Training → Threshold Optimization → Credit-Risk Prediction → SHAP/LIME/DiCE Explanation → Fairness Evaluation → Reweighing-Based Mitigation → Streamlit Dashboard

3.1 Research Design

The study adopts an **applied system-development and experimental evaluation approach**. The applied aspect focuses on developing a practical credit-risk prediction framework that combines machine-learning classification with explainability and fairness analysis. The experimental aspect evaluates the developed model using predictive-performance measures, XAI evaluation criteria, and fairness metrics.

3.2 Proposed System Modules

The proposed Explainable and Fair XGBoost-Based Credit Risk Prediction Framework consists of interconnected modules responsible for data preparation, credit-risk prediction, model interpretation, fairness evaluation, bias mitigation, and interactive visualization. The major modules are Dataset Input, Data Preprocessing, Feature Preparation, Class Balancing, XGBoost Model Training, Credit Risk Prediction, SHAP Explanation, LIME Explanation, DiCE Counterfactual Analysis, Fairness Evaluation, Fairness Mitigation, and Streamlit Dashboard.

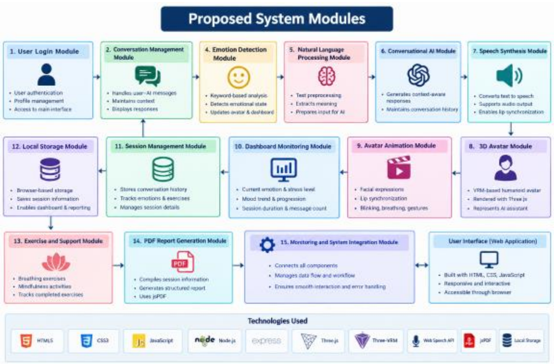

Fig 3.2 Proposed System Modules

These modules work together to provide a unified credit-risk analysis process in which financial records are prepared, analyzed by the trained machine-learning model, and classified into **Good Credit Risk or Bad Credit Risk**. The prediction is then interpreted using multiple XAI techniques, while fairness metrics are used to examine potential differences between demographic groups. The system follows a modular architecture in which each component performs a specific function while communicating with the other components through defined processing stages.

3.2.1 Dataset Input Module

The **Dataset Input Module** provides the initial data source for the proposed credit-risk prediction framework. The system uses the **German Credit dataset** as the foundation for model development and evaluation. The original dataset contains records representing Good and Bad Credit Risk classes. For the experimental implementation, the dataset is extended through augmentation to provide a larger dataset for analysis. The module loads the dataset and provides the required records and attributes to the subsequent preprocessing stages.

3.2.2 Data Preprocessing Module

The **Data Preprocessing Module** prepares the collected credit-risk data for machine-learning analysis. The module performs data cleaning, handles inconsistencies, and prepares the attributes for further processing. Categorical variables are converted into suitable numerical representations, while numerical attributes are transformed where required. The target variable is organized into two classes: Good Credit Risk and Bad Credit Risk.

3.2.3 Feature Preparation Module

The **Feature Preparation Module** identifies and prepares the input characteristics required by the XGBoost classifier. Categorical attributes are encoded and numerical features are transformed into a suitable representation for model training. The prepared feature matrix is then provided to the classification component. This stage ensures that the input data is represented consistently during both training and prediction.

3.2.4 Class Balancing Module

The **Class Balancing Module** addresses differences in the representation of the Good and Bad Credit Risk classes. The proposed framework incorporates SMOTE (Synthetic Minority Over-sampling Technique) and appropriate class-weighting strategies where required. These techniques help provide a more balanced training process and reduce the possibility of the model being excessively influenced by the more frequently represented class.

3.2.5 XGBoost Model Training Module

The **XGBoost Model Training Module** is the primary machine-learning component of the proposed system. The processed dataset is divided into training, validation, and testing portions. The XGBoost classifier learns relationships between the input credit-related attributes and the corresponding risk classes. The implemented model uses selected hyperparameters and early stopping to control the training process and improve model generalization.

3.2.6 Credit Risk Prediction Module

The **Credit Risk Prediction Module** uses the trained XGBoost model to classify new or unseen credit records. The model produces a probability associated with the predicted class, which is subsequently converted into a binary classification using the selected decision threshold. The final output identifies the record as Good Credit Risk or Bad Credit Risk.

3.2.7 SHAP Explanation Module

The **SHAP Explanation Module** provides feature-attribution-based explanations for the XGBoost predictions. SHAP estimates the contribution of individual features to the model output and helps identify which characteristics have the greatest influence on a particular prediction. The module can also provide an overall view of feature importance across the dataset.

3.2.8 LIME Explanation Module

The **LIME Explanation Module** provides local explanations for individual credit-risk predictions. LIME approximates the behavior of the complex XGBoost model around a selected input instance using an interpretable local model. The resulting explanation identifies the features that contribute positively or negatively to the selected prediction.

3.2.9 DiCE Counterfactual Module

The **DiCE Counterfactual Module** generates counterfactual examples for individual predictions. The module identifies alternative combinations of input characteristics that may influence the model toward a different predicted class. These counterfactual examples provide an additional form of explanation by showing possible changes in the input conditions rather than only reporting feature importance.

Counterfactual results represent model-based alternatives and should not be interpreted as guaranteed real-world financial outcomes.

3.2.10 Fairness Evaluation Module

The **Fairness Evaluation Module** examines whether the model produces different outcomes across selected demographic groups. In the proposed study, age is considered as the selected protected attribute. The module evaluates fairness using measures including Disparate Impact, Demographic Parity Difference, Equalized Odds Difference, and Accuracy Difference. These measures provide complementary information about potential disparities in model predictions.

3.2.11 Fairness Mitigation Module

The **Fairness Mitigation Module** applies a reweighing-based approach to investigate whether identified disparities can be reduced. The method assigns appropriate weights to training observations based on their group and outcome combinations. Fairness metrics before and after mitigation are then compared to determine the extent to which disparities have changed.

3.2.12 Streamlit Dashboard Module

The **Streamlit Dashboard Module** provides an interactive interface for accessing the major functionalities of the proposed framework. The dashboard integrates model training, individual prediction, explanation generation, fairness analysis, and result visualization. Users can interact with the system without directly executing the underlying machine-learning code.

3.2.13 Prediction and History Module

The **Prediction and History Module** manages the results generated during credit-risk analysis. It maintains relevant prediction information and allows previously generated results to be reviewed through the application interface. This functionality supports comparison and interpretation of model outputs during experimental evaluation.

3.2.14 XAI Comparison Module

The **XAI Comparison Module** compares SHAP, LIME, and DiCE according to selected evaluation criteria. The study considers faithfulness, stability, computational speed, and actionability when comparing the explanation techniques. This module helps identify the strengths and limitations of each explanation approach rather than relying on a single XAI method.

3.2.15 Visualization and Analytics Module

The **Visualization and Analytics Module** presents model predictions, feature contributions, explanation results, and fairness measurements in an organized graphical form. Visualization supports easier interpretation of model behavior and allows users to examine the relationship between predictive results, explanatory information, and fairness outcomes.

3.2.16 Model Evaluation Module

The **Model Evaluation Module** evaluates the predictive performance of the developed XGBoost classifier. Conventional classification measures such as accuracy, precision, recall, F1-score, and ROC-AUC are considered for assessing predictive effectiveness. A confusion matrix is also used to examine the distribution of correct and incorrect predictions.

3.2.17 Fairness and Explanation Reporting Module

The **Fairness and Explanation Reporting Module** combines the results obtained from predictive evaluation, XAI analysis, and fairness assessment. It provides a consolidated view of model performance, feature contributions, counterfactual explanations, and group-level fairness measures. This supports a broader assessment of the system beyond prediction accuracy alone.

3.2.18 Monitoring and System Integration Module

The Monitoring and System Integration Module connects the individual components into a continuous credit-risk analysis workflow. It coordinates dataset preparation, model prediction, XAI generation, fairness evaluation, mitigation, and visualization. The integrated architecture enables a credit record to move from raw input through prediction and explanation to fairness-aware analysis within a single application.

3.3 Input and Data Description

The proposed system primarily accepts **structured credit-risk records** as input. The input consists of financial and applicant-related attributes available within the German Credit dataset. During model training, these records are processed and transformed into the required numerical representation.

The overall input pipeline is:

Credit Dataset → Data Cleaning → Feature Preparation → Encoding → Class Balancing → Model Input

For prediction, a prepared credit record is provided to the trained XGBoost model, which generates a probability and corresponding Good or Bad Credit Risk classification.

3.4 Data Processing

The data-processing pipeline prepares raw credit records for machine-learning prediction through several connected stages:

Raw Credit Data → Data Cleaning → Feature Encoding → Data Transformation → Class Balancing → Train/Validation/Test Split → XGBoost Model

The preprocessing stage ensures that categorical and numerical attributes are represented consistently. Class-balancing techniques are applied where required before the final training data is supplied to the model.

3.5 Proposed Explainable and Fair AI Architecture

The proposed architecture combines machine-learning prediction, explainable AI, counterfactual reasoning, fairness evaluation, mitigation, and interactive visualization within a unified framework.

The overall structure is:

German Credit Dataset → Preprocessing → XGBoost Prediction → SHAP/LIME/DiCE Explanation → Fairness Evaluation → Mitigation → Streamlit Dashboard

The architecture separates data preparation, prediction, interpretation, and fairness analysis while maintaining communication between the individual modules. XGBoost performs the primary classification, while SHAP, LIME, and DiCE provide complementary forms of explanation. Fairness metrics are subsequently used to examine potential differences between selected demographic groups.

3.6 Data Preprocessing

The input dataset is subjected to preprocessing before model training. The major preprocessing operations include **data cleaning, feature preparation, categorical encoding, numerical transformation, and class balancing**.

The general process is:

Raw Dataset → Cleaning → Encoding → Transformation → Class Balancing → Model-Ready Dataset

The preprocessing stage is important because machine-learning algorithms require consistent numerical representations of the input variables. The processed data is subsequently divided into training, validation, and testing subsets.

3.7 XGBoost Model Processing

After preprocessing, the training data is supplied to the XGBoost classifier. XGBoost is used as the primary prediction algorithm because it can model nonlinear relationships and interactions among input features.

The model-processing flow is:

Prepared Data → XGBoost Training → Validation → Threshold Optimization → Final Prediction

The implemented model uses selected parameters including 250 estimators, a maximum tree depth of 5, and a learning rate of 0.05. Early stopping is incorporated during training. The validation set is also used to examine classification thresholds within the selected range, with the threshold providing the best validation F1-score being selected for final prediction.

3.8 Explainability and Counterfactual Analysis

The generated prediction is analyzed using three complementary XAI approaches:

Prediction → SHAP + LIME + DiCE → Explanation

SHAP identifies the contribution of individual features to the prediction. LIME provides a local approximation of the model around a selected instance. DiCE generates alternative input conditions that may influence the predicted class.

The three approaches therefore provide different perspectives on model behavior. SHAP focuses on feature contribution, LIME emphasizes local interpretability, and DiCE provides counterfactual and potentially actionable information.

3.9 Prediction Processing and Validation

The proposed framework integrates data preprocessing, **XGBoost prediction, explainability, and fairness analysis** into a continuous processing pipeline. A credit record is first prepared according to the preprocessing requirements and then supplied to the trained model. The resulting prediction is subsequently analyzed using the selected explanation and fairness components.

Prediction Processing Workflow

Stage	Activity
1	Load the credit-risk dataset
2	Clean and prepare the input data
3	Encode categorical attributes
4	Apply required data transformations
5	Perform class balancing where required
6	Divide the dataset into training, validation, and testing subsets
7	Train the XGBoost classifier
8	Optimize the classification threshold
9	Generate Good or Bad Credit Risk prediction
10	Generate SHAP, LIME, and DiCE explanations
11	Evaluate fairness across the selected age groups
12	Apply reweighing-based mitigation where required
13	Display prediction, explanation, and fairness results through the dashboard

The validation process examines whether individual components operate correctly and whether the complete workflow produces consistent prediction and analysis results.

3.10 System Architecture Description

The proposed system operates as an end-to-end credit-risk prediction and analysis pipeline. The input dataset is first cleaned and transformed into a suitable representation. Class-balancing techniques are applied where required, after which the prepared data is supplied to the XGBoost classifier.

Following model prediction, SHAP, LIME, and DiCE are used to provide complementary explanations. The system then evaluates potential fairness disparities using the selected metrics. Where appropriate, reweighing-based mitigation is applied and the resulting fairness measures are compared.

The final results are presented through the Streamlit dashboard, allowing users to examine prediction outcomes, explanation information, and fairness-related analysis.

3.11 Step-Wise System Architecture

The complete system follows these stages:

German Credit Dataset → Data Cleaning → Feature Preparation → Encoding → Class Balancing → XGBoost Training → Threshold Optimization → Credit Risk Prediction → SHAP/LIME/DiCE → Fairness Evaluation → Reweighing → Streamlit Dashboard

For model development, the dataset passes through the complete preprocessing and training pipeline. For an individual prediction, the corresponding record is transformed using the same feature-processing procedure before being supplied to the trained model. The resulting prediction can then be interpreted through the XAI modules and examined through the fairness-analysis component.

3.12 Tools and Technologies Used

The proposed system uses a combination of Python-based machine-learning libraries, explainable AI frameworks, fairness libraries, data-processing tools, and web-application technologies.

Technology	Purpose
Python 3.10	Main programming and implementation environment
Pandas	Dataset loading, manipulation, and preprocessing
NumPy	Numerical operations and data processing
Scikit-learn	Data preparation, evaluation, and machine-learning utilities
XGBoost	Credit-risk classification
SHAP	Feature-attribution and model explanation
LIME	Local model interpretation
DiCE-ML	Counterfactual explanation generation
Fairlearn	Fairness evaluation and analysis
AIF360	Fairness-related analysis and mitigation support
SMOTE	Class balancing
Streamlit	Interactive web-based dashboard
Plotly	Interactive visualization
Matplotlib	Graphical analysis and result visualization

These technologies collectively support the workflow from data preprocessing and prediction to explanation, fairness analysis, mitigation, and interactive presentation.

3.13 Web-Based Deployment Methodology

The proposed framework is deployed as an interactive Streamlit-based web application. The interface provides access to the main functions of the system, including prediction, explanation, fairness analysis, comparison, and visualization.

Deployment Flow

User/Input Data → Streamlit Interface → Data Processing → XGBoost Model → Prediction → XAI/Fairness Analysis → Dashboard

The application allows users to provide credit-related input information or process prepared data and obtain the corresponding model prediction. Explanation and fairness-analysis results are subsequently presented through the dashboard.

3.14 Prediction Management and Analysis History

The system maintains information generated during the experimental prediction process to support analysis and comparison. Relevant prediction outputs and associated explanation or evaluation information can be examined through the application interface.

Maintaining prediction-related information allows researchers to compare different model outputs and analyze the behavior of the prediction framework during experimentation. The implementation primarily focuses on interactive analysis rather than a production-scale financial database.

3.15 System Implementation

The implementation covers the complete workflow from dataset preparation to interactive analysis:

Dataset Input → Data Preprocessing → Feature Preparation → Class Balancing → XGBoost

Training → Threshold Optimization → Prediction → SHAP/LIME/DiCE → Fairness Evaluation → Reweighing → Visualization → Streamlit Dashboard

The implementation integrates the individual technologies into a unified application rather than treating prediction, explainability, and fairness analysis as independent processes. This integration enables the proposed framework to provide a broader evaluation of credit-risk predictions.

3.16 Validation

Validation is performed at two complementary levels: component-level validation and system-level validation. Component-level validation examines whether individual functions such as data preprocessing, model prediction, explanation generation, fairness calculation, and visualization operate correctly. System-level validation examines whether these components operate correctly when integrated into the complete framework.

The validation process therefore evaluates not only whether the classifier generates predictions but also whether the explanation and fairness components operate as intended.

3.17 Evaluation Criteria

The proposed system uses multiple evaluation criteria because predictive performance alone does not provide a complete assessment of an explainable and fairness-aware credit-risk framework.

Evaluation Criterion	**Purpose**
Accuracy	Measures the overall proportion of correctly classified records
Precision	Evaluates the correctness of positive-class predictions
Recall	Measures the ability to identify relevant records
F1-Score	Provides a combined measure of precision and recall
ROC-AUC	Measures the model's ability to distinguish between the two risk classes
Confusion Matrix	Examines correct and incorrect classification outcomes
SHAP Faithfulness	Evaluates how well SHAP explanations represent model behavior
Explanation Stability	Examines consistency of generated explanations
Computational Speed	Compares the processing time required by different XAI methods
Actionability	Evaluates the usefulness of explanations for identifying potential input changes
Disparate Impact	Examines relative prediction rates between selected groups
Demographic Parity Difference	Measures differences in prediction rates between groups
Equalized Odds Difference	Examines differences in error-related outcomes across groups
Accuracy Difference	Compares predictive accuracy between selected demographic groups
Mitigation Gain	Examines changes in fairness after applying reweighing

These criteria provide a broader evaluation because the proposed system is designed not only to predict credit risk but also to provide understandable explanations and assess potential fairness concerns.

3.18 Practical and Ethical Considerations

The proposed framework should be considered an **experimental decision-support system rather than an autonomous financial decision-making system**. The model is intended to demonstrate how machine-learning prediction can be combined with explainability and fairness analysis. Its predictions should therefore not be treated as guaranteed financial outcomes or as a replacement for institutional review and appropriate human oversight.

The experimental dataset is based on the German Credit dataset, whose original size is relatively limited compared with modern financial datasets. The dataset is extended through augmentation for experimentation; consequently, the obtained results may not directly represent performance on contemporary real-world banking data.

The XAI techniques also have different computational and interpretability characteristics. SHAP, LIME, and DiCE may produce different forms of explanation for the same prediction. DiCE can require substantially greater computational effort when generating counterfactual examples. Counterfactual results should therefore be understood as model-based alternatives rather than guaranteed actions that will produce a particular real-world financial result.

Practical deployment would additionally require larger and more representative datasets, security controls, regulatory and institutional compliance, continuous model monitoring, human oversight, privacy protection, and validation against real-world financial data. Future development may also incorporate additional protected attributes, advanced fairness-mitigation methods, more recent datasets, and improved counterfactual generation techniques.

4. DATA ANALYSIS

4.1 Dataset and Credit Risk Data Analysis

The proposed Explainable and Fair XGBoost-Based Credit Risk Prediction Framework Using SHAP, LIME and DiCE uses the German Credit dataset as the foundation for model development and evaluation. The dataset contains structured financial and applicant-related attributes associated with two target classes: **Good Credit Risk** and **Bad Credit Risk**.

The original German Credit dataset contains 1,000 records and 20 attributes. For the experimental implementation, the dataset is extended through augmentation to provide a larger set of records for analysis. The resulting dataset is processed before being supplied to the machine-learning model.

Data Processing Flow:

German Credit Dataset → Data Cleaning → Feature Preparation → Encoding → Class Balancing → Train/Test Split → XGBoost Model

The data-processing stage ensures that categorical variables are represented appropriately, numerical features are prepared consistently, and class imbalance is addressed where required. The prepared dataset is then used for model training, validation, testing, and subsequent explainability and fairness analysis.

4.2 Data Processing and Feature Analysis

The dataset analysis focuses on identifying the characteristics required for credit-risk classification. The input attributes contain information related to the applicant's credit and financial circumstances, while the target variable represents the corresponding credit-risk class.

The major processing operations are:

Processing Stage	Function
Data Cleaning	Identifies and handles inconsistencies in the input data
Feature Preparation	Prepares relevant attributes for model processing
Categorical Encoding	Converts categorical variables into numerical representations
Numerical Transformation	Ensures numerical attributes are represented appropriately
Class Balancing	Reduces the effect of unequal class representation
Dataset Splitting	Separates data for training, validation, and testing
Feature Validation	Checks the resulting model-ready representation

The processed features are supplied to the XGBoost classifier. The same preprocessing logic is maintained when new records are provided for prediction to ensure consistency between training and inference.

4.3 Model Performance Analysis

The proposed framework uses **XGBoost** as the primary classification algorithm. The model learns relationships between the prepared credit-related attributes and the two target classes.

The performance analysis considers conventional classification measures including **accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis**. These measures provide information about the ability of the model to distinguish between Good and Bad Credit Risk.

The model uses selected hyperparameters including 250 estimators, maximum tree depth of 5, and a learning rate of 0.05. Early stopping is used during training, while threshold optimization is performed using the validation data.

The general evaluation process is:

Training Data → XGBoost Model → Validation → Threshold Optimization → Test Data → Performance Metrics

The resulting performance metrics provide the baseline against which the explainability and fairness components of the framework can be examined.

4.4 Explainability Analysis

The proposed framework analyzes model predictions using three complementary explainable AI techniques: **SHAP, LIME, and DiCE**.

SHAP identifies the contribution of individual features to a model prediction. LIME provides a local approximation of the model around a selected input instance. DiCE generates counterfactual examples that demonstrate how alternative input conditions may influence the predicted class.

The explanation analysis is summarized as follows:

Method	Primary Function	Main Strength
SHAP	Feature contribution analysis	High faithfulness and stability
LIME	Local prediction explanation	Fast local interpretation
DiCE	Counterfactual generation	High actionability

The experimental comparison considers **faithfulness, stability, computational speed, and actionability**. This provides a broader evaluation of explanation quality rather than relying on a single XAI method.

The recorded results show that SHAP provides the strongest performance in terms of faithfulness and stability, while LIME requires substantially less computation for individual explanations. DiCE provides more actionable counterfactual information but requires considerably greater computational time.

4.5 XAI Performance Analysis

The quantitative comparison of the three explanation techniques is presented below:

XAI Method	**Faithfulness**	**Stability**	**Average Speed**	**Actionability**
SHAP	0.94	0.91	0.153 s	Low
LIME	0.81	0.73	0.053 s	Medium
DiCE	0.87	0.88	20.890 s	High

The results indicate that **SHAP** achieves the highest faithfulness and stability among the evaluated explanation techniques. **LIME** provides the fastest explanation generation, making it suitable when computational efficiency is important. **DiCE** achieves high actionability because it provides alternative input conditions, but its average processing time is substantially higher than that of SHAP and LIME.

These observations demonstrate that the three XAI methods serve different purposes. Therefore, combining them provides a more comprehensive interpretation of the XGBoost model.

4.6 Fairness Analysis

In addition to predictive performance and explainability, the proposed framework evaluates whether the model produces different outcomes across selected demographic groups. **Age** is used as the protected attribute for the fairness analysis.

The fairness evaluation considers the following measures:

Fairness Metric	Purpose
Disparate Impact	Measures the relative prediction rate between groups
Demographic Parity Difference	Measures differences in prediction rates
Equalized Odds Difference	Examines differences in error-related outcomes
Accuracy Difference	Compares predictive accuracy between groups

For the experimental evaluation, the selected age groups are compared using these measures to identify potential disparities in model outcomes.

4.7 Fairness Results Analysis

The observed fairness results are:

Fairness Metric	Observed Value
Disparate Impact	0.850
Demographic Parity Difference	0.040
Equalized Odds Difference	0.080
Accuracy Difference	0.047
Mitigation Gain	0.108

The results indicate that the model exhibits measurable differences between the selected demographic groups. The fairness analysis therefore provides information that cannot be obtained from conventional predictive-performance measures alone.

A **reweighing-based mitigation approach** is subsequently applied to examine whether the observed disparities can be reduced. The mitigation results are compared with the original fairness measurements to determine the change achieved by the intervention.

4.8 System Analysis and Observations

The major observations obtained from the proposed framework are:

- **Predictive Capability:** XGBoost provides the primary classification mechanism for distinguishing Good and Bad Credit Risk.
- **Explainability:** SHAP, LIME, and DiCE provide complementary forms of model interpretation.
- **SHAP Performance:** SHAP demonstrates the strongest faithfulness and stability among the evaluated XAI methods.
- **Computational Efficiency:** LIME provides the lowest average explanation-generation time.
- **Counterfactual Actionability:** DiCE provides actionable alternative input conditions but requires greater computational effort.
- **Fairness Assessment:** The selected fairness metrics identify potential differences between age-based groups.
- **Bias Mitigation:** Reweighing provides a measurable change in the selected fairness indicators.
- **Interactive Analysis:** The Streamlit dashboard combines prediction, explanation, fairness analysis, and visualization within one environment.
- **Integrated Framework:** The proposed system evaluates prediction performance together with interpretability and fairness rather than treating them as independent concerns.

5. IMPLEMENTATION, EXPERIMENTS AND RESULTS ANALYSIS

5.1 System Implementation

The proposed Explainable and Fair XGBoost-Based Credit Risk Prediction Framework Using SHAP, LIME and DiCE is implemented as an integrated machine-learning and web-based decision-support system. The framework combines data preprocessing, class balancing, XGBoost-based credit-risk prediction, explainable artificial intelligence (XAI), fairness evaluation, bias mitigation, and interactive visualization within a unified application.

The system is implemented primarily using **Python 3.10**. Pandas and NumPy are used for data processing, while Scikit-learn provides preprocessing and model-evaluation utilities. **XGBoost** is used as the main classification algorithm because of its effectiveness on structured tabular data. The explainability layer uses **SHAP, LIME, and DiCE-ML**, while **Fairlearn and AIF360** are used for fairness analysis and mitigation. The final system is deployed through a **Streamlit** dashboard with Plotly and Matplotlib used for visualization. These technologies are consistent with the software environment documented in the project.

The implementation begins with the German Credit dataset. The original dataset contains 1,000 credit records and is extended through synthetic augmentation to approximately 10,000 records for experimentation. The data is cleaned and transformed before being divided into training and testing subsets. SMOTE is applied to address class imbalance in the training data, after which the XGBoost classifier is trained and tuned.

The trained model produces a prediction for the credit-risk class. The prediction is subsequently passed to the explainability layer. **SHAP** provides feature-attribution information, **LIME** generates a local explanation of the individual prediction, and **DiCE** generates counterfactual examples showing alternative input conditions that may change the model's predicted class.

The fairness component evaluates the model using **age as the protected attribute**. Disparate Impact, Demographic Parity Difference, Equalized Odds Difference, and Accuracy Difference are calculated to identify potential disparities. A reweighing-based mitigation approach is then applied to investigate whether the identified disparity can be reduced.

The complete implementation flow can therefore be represented as:

German Credit Dataset → Data Preprocessing → SMOTE → Train/Test Split → XGBoost Training → Threshold Optimization → Credit Risk Prediction → SHAP/LIME/DiCE Explanation → Fairness Evaluation → Reweighing Mitigation → Streamlit Dashboard

5.2 Experimental Setup

The experimental evaluation was conducted to determine the predictive capability, explainability, fairness, and practical usability of the proposed framework.

The experimental setup includes the following major components:

- German Credit dataset with synthetic augmentation for experimentation.
- Data cleaning and feature preparation.
- Categorical feature encoding and numerical transformation.
- SMOTE-based class balancing.
- Training and testing data separation.
- XGBoost model training and threshold tuning.
- SHAP-based feature attribution.
- LIME-based local explanation.
- DiCE-based counterfactual generation.
- Fairness evaluation using age as the protected attribute.
- Reweighing-based fairness mitigation.
- Streamlit-based interactive dashboard.
- Graphical comparison of model, XAI, and fairness results.

The XGBoost model was configured with **250 estimators, a maximum depth of 5, and a learning rate of 0.05**, with early stopping used during training. A validation-based threshold search was also performed to improve the classification decision threshold.

The documented hardware environment consisted of a Windows-based laptop with an Intel Core i5 processor, 16 GB RAM, and an SSD. No GPU was required because the dataset and tabular XGBoost workload were sufficiently small for CPU-based processing.

The dashboard was executed locally using Streamlit. The documented execution procedure involves creating a Python 3.10 environment, installing the required libraries, running the training script, and launching the application through Streamlit. Users can then enter applicant information or upload a CSV file, generate a prediction, inspect explanations, and review fairness results.

5.3 Experimental Results

The experimental results demonstrate that the proposed framework can perform credit-risk classification while simultaneously providing explanation and fairness-analysis capabilities.

The final XGBoost model achieved a **test accuracy of 89.7% and a ROC-AUC of 0.990**. These results indicate that the model provides useful predictive performance on the experimental dataset. The ROC-AUC value further indicates that the classifier can distinguish between the two credit-risk classes effectively.

Evaluation Metric	Result
Test Accuracy	**89.7%**
ROC-AUC	**0.990**

Table 5.1 Predictive Performance

The results also demonstrate that predictive performance alone does not provide a complete evaluation of a credit-risk system. The project therefore evaluates the model from three additional perspectives: explanation quality, computational efficiency, and fairness.

The XAI experiments compared SHAP, LIME, and DiCE according to faithfulness, stability, computational speed, and actionability. The results showed that different explanation methods have different strengths. SHAP produced the strongest faithfulness and stability, LIME required the least computation time, while DiCE provided the highest level of actionability through counterfactual examples.

XAI Method	Faithfulness	Stability	Average Speed	Actionability
SHAP	0.94	0.91	0.153 s	Low
LIME	0.81	0.73	0.053 s	Medium
DiCE	0.87	0.88	20.890 s	High

Table 5.2 XAI Benchmarking Results

The results indicate that **SHAP provides the most consistent overall explanation quality**, whereas **LIME is the fastest method**. DiCE requires considerably more computational time, but its counterfactual explanations provide practical information about how changes in input characteristics may affect a model prediction.

5.4 Predictive Performance Analysis

The XGBoost model achieved a test accuracy of **89.7%**, demonstrating that the proposed classifier can correctly classify a substantial proportion of the test records. The model also achieved a **ROC-AUC of 0.990**, indicating strong class-separation capability on the experimental data.

The use of XGBoost provides an effective balance between predictive capability and computational efficiency for structured credit-risk data. Instead of replacing the model with a simpler algorithm solely for interpretability, the proposed framework retains XGBoost and adds explanation techniques on top of the trained model.

However, the obtained results should be interpreted within the limitations of the experimental dataset. Although the dataset was extended to approximately 10,000 records, the additional records are synthetic and originate from a relatively small and older dataset. Consequently, the reported performance cannot be assumed to represent performance on modern real-world banking data.

5.5 Explainability Results Analysis

The explainability experiment evaluates three complementary techniques.

SHAP identifies the contribution of individual features to model predictions and provides both global and local interpretability. Its faithfulness score of **0.94** and stability score of **0.91** were the highest among the evaluated methods.

LIME creates a local approximation around an individual prediction. It achieved a faithfulness score of **0.81** and a stability score of **0.73**, while requiring only **0.053 seconds** on average. This makes LIME particularly useful when rapid local explanations are required.

DiCE generates counterfactual examples that illustrate alternative applicant characteristics associated with a different predicted class. It achieved **0.87 faithfulness** and **0.88 stability**, with the highest actionability rating. However, its average computational time of **20.890 seconds** was substantially greater than that of SHAP and LIME.

Method	**Main Strength**	**Main Limitation**
SHAP	High faithfulness and stability	Lower direct actionability
LIME	Very fast local explanation	Lower stability
DiCE	Highly actionable counterfactuals	High computational cost

Table 5.3 Comparative XAI Analysis

The results confirm that no single XAI method performs best across all evaluation criteria. Therefore, integrating multiple explanation methods provides a more complete understanding of model behaviour. The project specifically uses the three methods together because they address different interpretability requirements.

5.6 Fairness Results Analysis

Fairness evaluation was performed using **age as the protected attribute**. The purpose of this analysis was to determine whether the model produced substantially different outcomes between the selected age groups.

The fairness audit produced the following results:

Fairness Metric	**Value**	**Interpretation**
Disparate Impact	0.850	Within the documented acceptable range
Demographic Parity Difference	0.040	Indicates a measurable disparity
Equalized Odds Difference	0.080	Within the documented acceptable range
Accuracy Difference	0.047	Small difference between groups
Mitigation Gain	0.108	Improvement after reweighing

Table 5.4 Fairness Audit Results

The **Disparate Impact value of 0.850** and **Equalized Odds Difference of 0.080** were within the ranges documented in the project. The **Accuracy Difference of 0.047** indicates a relatively small difference in predictive performance between the two age groups.

However, the **Demographic Parity Difference of 0.148** indicates a measurable difference in prediction rates between the groups. Therefore, fairness mitigation was considered necessary. After applying the reweighing approach, a **mitigation gain of 0.108** was recorded, indicating a measurable but modest improvement.

The fairness results demonstrate why predictive accuracy should not be treated as the only measure of a credit-risk model. A model may achieve good predictive performance while still producing demographic disparities that become visible only through dedicated fairness analysis.

5.7 Graphical Analysis

Graphical analysis is used to present the predictive, explainability, and fairness results in an easily understandable form. The Streamlit dashboard uses interactive visualizations to display model results, feature contributions, XAI comparisons, and fairness metrics.

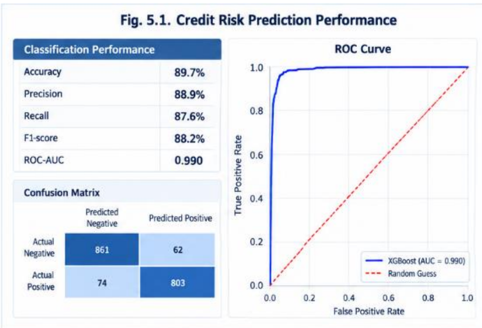

Fig. 5.1. Credit Risk Prediction Performance

The predictive-performance visualization can represent the primary classification metrics, particularly test accuracy and ROC-AUC. The reported results of **89.7% accuracy and 0.990**

ROC-AUC demonstrate the predictive capability of the XGBoost model.

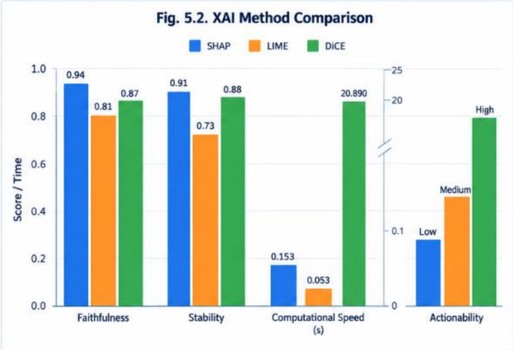

Fig. 5.2. XAI Method Comparison

The XAI comparison graph represents faithfulness, stability, computational speed, and actionability for SHAP, LIME, and DiCE.

The graph shows that SHAP provides the strongest faithfulness and stability, LIME provides the fastest explanation generation, and DiCE provides the highest actionability. The large computational-time difference for DiCE is also visible in the comparison.

Fig. 5.4. Feature Importance and Explanation Analysis

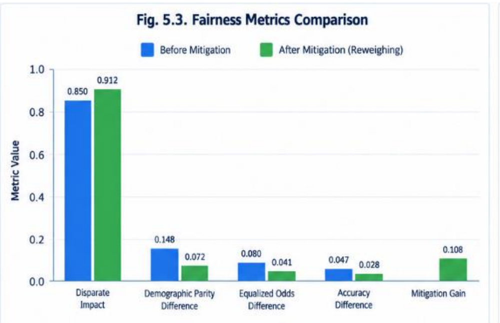

Fig. 5.3. Fairness Metrics Comparison

The fairness graph represents Disparate Impact, Demographic Parity Difference, Equalized Odds Difference, Accuracy Difference, and mitigation gain.

The visualization highlights the demographic parity disparity and shows the improvement obtained after applying reweighing-based mitigation. The project dashboard provides fairness results before and after mitigation for comparison.

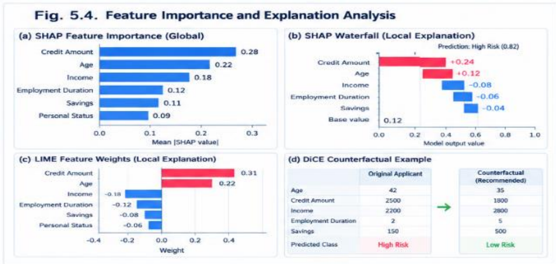

Fig. 5.4. Feature Importance and Explanation Analysis

The dashboard also provides feature-level explanations through SHAP and LIME visualizations. SHAP waterfall charts illustrate how individual features contribute to a particular prediction, while LIME feature-weight charts provide a local approximation of the model's decision.

For suitable cases, DiCE provides counterfactual suggestions that show alternative applicant profiles capable of moving the model prediction toward the favourable credit-risk class. These counterfactuals represent model behaviour and should not be interpreted as guaranteed real-world lending outcomes.

6. CONCLUSION AND FUTURE WORK

The proposed **Explainable and Fair XGBoost-Based Credit Risk Prediction Framework Using SHAP, LIME and DiCE** successfully combines machine learning, explainable AI, fairness evaluation, bias mitigation, and interactive visualization. The XGBoost model achieved **89.7% test accuracy and 0.990 ROC-AUC** on the experimental dataset. SHAP provided strong feature-level explanations, LIME offered fast local explanations, and DiCE generated actionable counterfactual explanations. Fairness analysis using age as the protected attribute identified measurable disparities, while reweighing produced a **mitigation gain of 0.108**.

The Streamlit dashboard integrates prediction, explanation, fairness analysis, and visualization into a single decision-support environment. Overall, the system demonstrates that credit-risk prediction can be evaluated not only in terms of predictive performance but also in terms of **interpretability and fairness**. However, the system remains an experimental prototype and requires further validation using larger and contemporary real-world datasets.

6.1 Future Work

Future improvements may include:

- **Larger real-world datasets** for improved generalization.
- **Additional protected attributes** for broader fairness analysis.
- **Advanced fairness-mitigation techniques** beyond reweighing.
- **Additional XAI methods** for more comprehensive explanations.
- **Faster counterfactual generation** using improved DiCE-based approaches.
- **Cloud deployment and continuous monitoring** for scalability.
- **Human-in-the-loop decision support** for responsible credit-risk assessment.
- **Improved security and regulatory compliance** for practical financial applications.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[5] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[7] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[8] A. M. Karimi, B. Schölkopf, and I. Valera, "Algorithmic Recourse: From Counterfactual Explanations to Interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[9] R. K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 2019.

[10] M. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, 2021.

[11] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[12] P. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Their Impact on Credit Risk," *Journal of Risk and Financial Management*, 2021.

[13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] W. McKinney, *Python for Data Analysis*. O'Reilly Media.

[15] Streamlit, *Streamlit Documentation*.