JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.9
eISSN:2805-5160

Counterfactual Pattern Mining for Interpretable Decision Support in Classification Systems

Julius Cesar O. Mamaril[1]*, Zuli Maulidati[2], Clark Kim C. Castro[3]

[1]ICT Management Office, Pangasinan State University, Philippines
[2]Information Systems, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[3]ICT Management Office, Pangasinan State University, Philippines

***Email:** jmamaril@psu.edu.ph[1]*, zulimaulidati@itats.ac.id[2], ckcastro_ms@psu.edu.ph[3]

Abstract

Machine learning classification models have achieved strong predictive performance across various decision-making domains, but understanding how predictions can change under alternative conditions remains a major challenge. This study proposes a counterfactual pattern mining framework for interpretable decision support in classification systems. The framework integrates counterfactual explanation generation with association pattern mining to identify minimal and recurring feature changes that may alter classification outcomes. Benchmark datasets from healthcare, finance, and marketing domains were used to evaluate the proposed approach. Classification performance was assessed using accuracy, precision, recall, F1-score, and AUC, while explanation quality was evaluated using validity, sparsity, plausibility, and computational efficiency. The benchmark evaluation results showed that the best average F1-score reached 0.86, while the counterfactual explanations achieved an average validity of 0.92, average sparsity of 2.47 changed features, and average plausibility of 0.81. These findings indicate that the proposed framework can generate concise, realistic, and outcome-changing explanations. By transforming individual counterfactual explanations into recurring patterns, this study contributes to explainable data mining and provides interpretable decision alternatives for classification-based decision support systems.

Keywords

Process innovation; Counterfactual explanation; Explainable AI; Pattern mining; Classification; Decision support

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026

Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Introduction

Background of Interpretable Classification Systems

There is a growing reliance on machine learning classification models for supporting decision-making across domains such as healthcare, finance, and marketing. In classification model healthcare applications, models are used to detect risk groups for patients, support the development of advanced alerting systems, and construct data-informed clinical decision support systems. In finance, classification models are used to evaluate credit requests, predict loan defaults, identify fraudulent activities, and evaluate finance-related risks. In marketing, classification models are used to forecast responses to marketing activities, evaluate the potential success of marketing activities, predict customer attrition, and analyze customers' buying activities. All of these applications demonstrate that classification models are used for more than just supporting prediction tasks of a technical nature (Ali et al., 2023). Classification models are applied to support and influence decisions of a practical nature.

Although classification models are valuable, the lack of interpretability, particularly with complex models, is a significant barrier. Complex but interpretable relationships with respect to a target variable can be modeled with tools such as ensemble classifiers and gradient boosting models. However, users still lack insight into the logic of the decisions these tools make. In many applied decision-support systems, the ability to accurately predict a target variable is not enough to support the decision. Stakeholders support the decision when the rationale for the prediction is clear and when the prediction is shown to be sensitive to changes in the decision variables (Hassija et al., 2023).

This issue becomes pronounced at the level of high-stakes prediction, as the acceptability of predictions hinges on the trust clinicians have in their transparency. Alqudah et al. (2026) noted that machine-learning techniques that utilize complex architectures, such as artificial neural networks and convolutional neural networks, are perceived as black-box techniques because they are not easily interpretable to the end-user. Given the constraints of transparency, clinicians may be reluctant to trust a model that classifies a case into a given diagnostic category, or that directs a case toward a specific recommendation. Therefore, there is a clear need for interpretable AI methods that rationalize model behavior and also provide interpretable recommendation strategies.

This concern is especially important in high-stakes situations where model outputs influence important decisions about people. A healthcare clinician may need to understand how certain risk factors contribute to an elevated risk classification. A financial analyst may need to explain why a credit application is classified as rejected. A marketing manager may need to know which customer behaviour changes are associated with a positive campaign response. In these cases, users are not only interested in the question of why a model made a prediction, but also in the question of what changes would be required to obtain a different prediction.

Therefore, interpretable classification systems should provide explanations that are understandable, relevant, and actionable. An explanation is more useful when it can guide decision-makers toward alternative scenarios rather than only describing model behaviour. This motivates the use of counterfactual explanations, which focus on identifying minimal changes that can alter the classification outcome.

Limitation of Conventional Explanation Methods

Several explainable artificial intelligence methods have been developed to improve the transparency of machine learning classification systems. Feature importance methods identify which variables contribute most strongly to model predictions. Local explanation methods such as SHAP and LIME estimate the contribution of each feature to an individual prediction. Explanations based on rules specify a formulation that is understandable to a human decision-maker and shows the reasoning for a given output. Such methods are useful to the extent that they provide insight to the decision-maker for the relevant factors that influence model behavior.

Use of conventional methods to provide explanations for a decision does not accommodate the rationale for the decision. For example, knowing the importance of a factor does not provide insight into the value of the factor needed to change the prediction. Although local explanations may specify the importance of a factor, the information may not be interpretable to the end-user. Relative to other explanation methods, explanations based on rules may come closest to identifying the required measure to change a given classification.

Standard methods for explanation provide some justification context around prediction-based decisions, but rarely incorporate actionable substitution scenarios that allow for the prediction to be achieved under different circumstances (Stepin et al., 2021). This limitation reduces their usefulness in decision-making contexts where users need to explore possible scenarios. For example, knowing that debt ratio is an important feature in a credit approval model is useful, but it is more actionable to know that changing debt ratio from high to medium, together with improving credit history from poor to good, may change the predicted outcome from rejected to approved. This type of explanation requires a counterfactual perspective.

Counterfactual Explanation as Actionable Interpretation

Counterfactual explanation provides an actionable form of interpretation by answering the question: what minimal changes would alter the classification outcome? Instead of only describing the importance of features, counterfactual explanation identifies an alternative version of an input instance that would lead the model to produce a different prediction. In this way, counterfactual explanation connects predictive modelling with decision alternatives.

For example, in a finance classification task, a model may classify a credit application as rejected. A counterfactual explanation may indicate that if the debt ratio changes from high to medium and credit history changes from poor to good, the prediction may change from rejected to approved. While feature importance measures generally rank actions according to their importance for a class prediction, in this case, an explanation offers a scenario in which the feature value is different and thus is more practical. In healthcare, a counterfactual explanation may show that changes in controllable indicators such as glucose level, body mass index, or blood pressure are associated with a lower predicted risk (Guidotti, 2022). In marketing, a counterfactual explanation may show that increasing customer engagement or improving previous campaign response may change the predicted outcome from no response to response.

The strength of counterfactual explanation lies in its ability to produce simple and outcome-oriented alternatives. A good counterfactual explanation should be valid, meaning that it successfully changes the predicted outcome. It should also be sparse, meaning that it requires only a small number of feature changes. In addition, it should be plausible, meaning that the suggested changes remain realistic according to the data distribution and domain constraints. These qualities make counterfactual explanation suitable for interpretable decision support.

However, counterfactual explanations are often generated at the individual instance level. When many counterfactual explanations are produced across a dataset, the results can become difficult to summarize (Bubeck et al., 2023). Decision-makers may need to identify not only individual alternatives, but also recurring patterns that appear across multiple cases. This creates the need to integrate counterfactual explanation with pattern mining.

Need for Counterfactual Pattern Mining

Limitations of counterfactual explanation include the lack of actionable support beyond a single prediction. From a practical standpoint, decision-makers generally do not focus on single-case decisions. If counterfactuals are generated for each case, the outputs will be overly granular. The outputs will be so numerous that it will be impossible to discern general principles, frequent modifications, or standard alternatives.

A solution to the previously stated limitation is the application of pattern mining to a collection of counterfactual explanations. Within the context of counterfactual explanation, the pattern type of interest is the association type, which identifies the most common subsets of alterations. If feature changes within a counterfactual explanation are represented in a transaction format, then each change can be represented as an item. In a counterfactual transaction, the change “debt ratio: high to medium” and “credit history: poor to good” can be represented as items. Mining these transactions will allow for the identification of frequent patterns of changes.

Pattern mining and counterfactuals lead to an integrated and more robust explanation framework. Counterfactuals focus on changes that impact predictions, while pattern mining describes these changes in consistent, interpretable patterns. Consequently, the output generated is more amenable to decision support, as users can identify frequent combinations of changes that lead to the desired classification. This framework deals with explanations that operate at the local level, and offers a pattern perspective on how changes to a set of instances may impact a given prediction.

Research Gap

Although explainable AI is getting a lot of attention, especially for classification systems, there are still a lot of challenges when it comes to designing explanations that are actionable and based on patterns. To begin with, most of the existing methods focus on local explanations and do not offer verdicts or decisions that are alternatives for multiple instances. Additionally, a lot of the individual counterfactuals created are not meant to be summarized, especially in the context of big data. Moreover, there are not many efforts focused on using association pattern mining to analyze counterfactual changes and how they affect classification results (Aldoseri et al., 2023). Finally, to create a really good decision support system, explanations ought to be valid, concise, and plausible… and while these qualities are not systematically evaluated in existing explanations, they are of utmost importance.

These gaps indicate the need for a framework that can transform individual counterfactual explanations into recurring patterns (Ali et al., 2023). Such a framework can help users understand not only what changes may alter a single prediction, but also which changes frequently appear across many cases. This study addresses this need by proposing a counterfactual pattern mining framework for interpretable decision support in classification systems.

Research Objectives

The main objective of this study is to develop a counterfactual pattern mining framework for interpretable decision support in classification systems. Specifically, this study aims to achieve the following objectives:

1. To develop a counterfactual pattern mining framework for interpretable decision support in classification systems.
2. To transform individual counterfactual explanations into structured counterfactual transactions.
3. To apply association pattern mining to identify recurring counterfactual patterns.
4. To evaluate the framework using healthcare, finance, and marketing benchmark datasets.
5. To assess the generated explanations using validity, sparsity, plausibility, and computational efficiency.

Research Contributions

This study makes several contributions to explainable data mining and interpretable classification-based decision support. First, it integrates counterfactual explanation with association pattern mining to produce explanation outputs that are both actionable and pattern-based. Second, it introduces a transaction-based representation of counterfactual changes, allowing individual counterfactual explanations to be transformed into mineable structures. Third, it provides recurring counterfactual patterns as interpretable decision alternatives, enabling users to identify common combinations of feature changes associated with altered classification outcomes (Dwivedi et al., 2022). Fourth, it evaluates explanation quality across healthcare, finance, and marketing benchmark domains using validity, sparsity, plausibility, and computational efficiency. Finally, it contributes to transparent decision support by connecting predictive classification outcomes with understandable and reusable explanation patterns. Figure 1 presents the proposed counterfactual pattern mining framework used in this study. The framework shows how classification outputs are transformed into counterfactual explanations and then converted into recurring patterns for decision support.

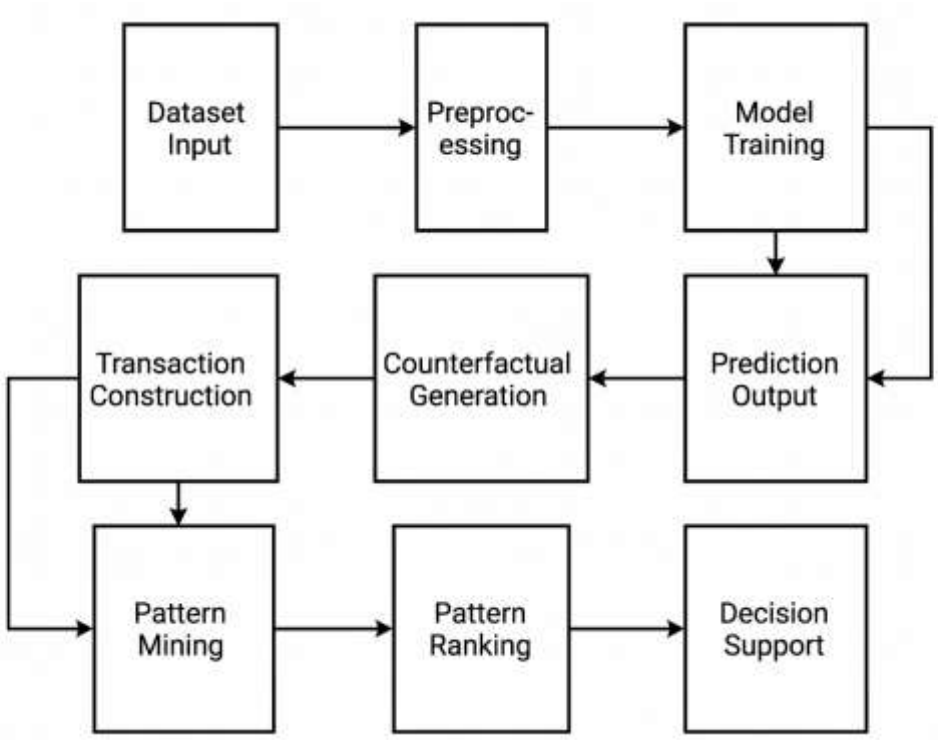

Figure 1. The proposed counterfactual pattern mining framework

As shown in Figure 1, the proposed framework does not stop at individual counterfactual explanations (Guidotti, 2022). Instead, it converts counterfactual changes into structured transactions so that recurring decision alternatives can be discovered through pattern mining.

Methodology

Research Design

This study utilized an experimental quantitative data science approach to develop and examine a counterfactual pattern mining framework designed to provide interpretable decision support within classification systems. This research was based on a computational experiment where benchmark datasets from health care, finance, and marketing decision-making were used. These domains were purposely chosen based on the frequent use of classification-based decisions in these areas. Decisions in these domains require not only classification and predictive accuracy but also the need for an interpretable explanation.

The methodological process involved several interrelated steps. First, benchmark datasets were gathered and modified in order to make the data amenable to both classification modeling and counterfactual analysis. Next, classification models were fit to the data in order to yield prediction results (Hollmann et al., 2025). Then, counterfactual explanations were developed in response to the predictions of the classification models. Following this, the counterfactual changes were operationalized as transactional items to be amenable to treatment in the association pattern mining framework. Subsequently, frequent counterfactual patterns were generated and, in the context of the decision support system, ranked for interpretability. In the final step, the framework was subjected to evaluation based on classification and explanation quality.

This design enabled the research to explore the dimensions of both prediction and explanation offered by the framework (Hassija et al., 2023). In the context of the classification of the framework, classification quality measures were evaluated, and for the explanation dimension, the quality of the counterfactual patterns was assessed based on the criteria of validity, sparseness, plausibility, and computational efficiency.

Benchmark Dataset Description

The experimental evaluation used controlled benchmark datasets representing healthcare, finance, and marketing decision-making contexts. These datasets provided controlled evaluation settings for examining the framework across different classification and decision-support contexts. The benchmark datasets used to assess the counterfactual pattern mining framework for different decision-making contexts have been compiled in Table 1.

Table 1 Summary of benchmark datasets

Domain	Dataset Description	Instances	Features	Target Class	Task Type
Healthcare	Clinical risk classification	768	8	High-risk / Low-risk	Binary classification
Finance	Credit approval classification	1,000	20	Approved / Rejected	Binary classification
Marketing	Campaign response classification	4,521	16	Respond / Not respond	Binary classification

The datasets used represent diverse feature and class distributions and decision-making support requirements. The healthcare dataset provides a clinical risk evaluation task that assesses case differentiation as high risk or low risk. The finance dataset corresponds to the case of credit evaluation, classifying the decision of case approval or rejection. The marketing dataset corresponds to the evaluation of a campaign, classifying the decision of case participation or non-participation.

The selected datasets were used to assess the ability of the proposed framework to produce interpretable counterfactual patterns across different classification contexts. Given the study's emphasis on the quality of explanation, the datasets were designed to facilitate counterfactual pattern extraction as well as predictive modeling.

Data Preprocessing

Data preprocessing was completed prior to the development of the classification models and the generation of counterfactual explanations. The goal of preprocessing was to improve the quality of the data and ensure consistency among the features. The datasets were made ready to undergo both machine learning and pattern mining analyses. Numerical data were made complete using the median since it is the least affected by outliers. Missing categorical data were made complete by using the most frequent category, or mode, in that feature.

To enable the classification models to process the data, categorical features were transformed using one-hot encoding. For the classification models to be able to process data meaningfully, numerical data were made to fall within a comparable range by the min-max normalization method. This was particularly important for the counterfactual generation process since the goal is to find the distance between the original instance and the counterfactual instance. If the goal is to find the distance, features with the largest ranges would obscure the purpose.

Each dataset was subjected to an 80-20 split, where 80% of the data was set aside for training, and 20% was set aside for testing. The training data were used to build the classification models, while the testing data were used to evaluate the models and generate counterfactuals. All continuous features were converted to discrete features using interpretable levels such as low, medium, and high. This was a necessary step in the process to accommodate the requirements of association pattern mining. For instance, a glucose value may be categorized as normal, moderate, or high, a debt ratio may be categorized as low, medium, or high, and an engagement score may be categorized as low, medium, or high.

Discretized feature values facilitated the formation of counterfactual transactions. This approach made it possible to encode the changes of interest in a more interpretable manner, such as “glucose: high to moderate” or “engagement: low to high.” Therefore, preprocessing was useful in the classification and the generation of logical counterfactual patterns.

Classification Model Development

Following preprocessing, the focus was on the development of classification models to generate prediction outputs for the benchmark datasets. This study utilized three classification models: Logistic Regression, Random Forest, and XGBoost. These models were chosen to align with the goals of the study and represented different model complexity levels and the respective classification capabilities.

Table 2 presents the classification models used to generate prediction outcomes before counterfactual explanations were produced.

Table 2 Classification models and experimental configuration

Model	Role	Main Setting	Reason for Use
Logistic Regression	Interpretable baseline	L2 regularization	Provides a simple and transparent baseline
Random Forest	Main classification model	100 trees, tuned max depth	Captures nonlinear feature interactions
XGBoost	Strong predictive baseline	Learning rate 0.05, tuned depth	Provides strong performance for tabular classification

The inclusion of different complexity levels in modeling ensures that this evaluation tests if the proposed framework is applicable to both interpretable and complex models. Logistic Regression represents the simple interpretable model, while Random Forest and XGBoost are capable of complex feature interactions. The assessment of these models enabled comparison of whether differences in model complexity influenced the generation of counterfactual explanations.

Counterfactual explanations were generated after the trained models were assessed for classification performance to ensure the hypotheses related to the framework were validated. This was important because poor classification performance is directly related to unexplanatory

counterfactuals. The classification models were assessed, and the best model was chosen for the generation of counterfactual explanations and the pattern mining analysis.

Counterfactual Explanation Generation

After the classification models generated predictions, counterfactual explanation generation was performed to find alternative examples that were similar to the original examples but resulted in different classification outcomes (Lee et al., 2026). Counterfactual explanations identify the minimum set of feature changes required to alter a prediction from its current class to a specified target class.

The counterfactual generation process was guided by the objective function presented in Equation 1.

$$CF(x) = argminx^{'}\left[d\left(x, x^{'}\right) + \lambda \cdot L\left(f\left(x^{'}\right), y_{t}arget\right)\right] \quad (1)$$

Given the original instance x, its counterfactual instance x', the distance between x and x' d(x, x'), the model prediction for the counterfactual instance f(x'), the target class y_target, and a trade-off parameter λ.

Counterfactual explanations aim to modify the prediction to the target class while remaining close to the original instance. The counterfactual distance encourages the generated counterfactual to remain similar to the original instance, while the prediction loss implies that the counterfactual should move towards the target classification. The balance is key because a good counterfactual should not suggest excessive or implausible changes (Ali et al., 2023).

This research evaluated counterfactual explanations in terms of minimality, actionability, stability, diversity, discriminative power, and running time (Guidotti, 2022). Validity examines whether a generated counterfactual changes the prediction, while sparsity is the number of features that need to be modified to achieve the counterfactual. Finally, plausibility examines whether the generated counterfactual remains realistic with respect to the observed benchmark data distribution and domain constraints (Stepin et al., 2021). These criteria help ensure that the counterfactual explanations are interpretable and valid.

Counterfactual Transaction Construction

The initial stage applied pattern mining on transaction-based representations of generated counterfactual explanations. This is because, in the context of the proposed method, counterfactual transactions are formed by treating each change of feature in a counterfactual explanation as an item. Thus, for the association pattern mining technique to be applied, data should be represented as a collection of items.

Consider a finance-related classification task. For instance, let the original prediction classify a certain credit application as rejected, while the counterfactual prediction classifies the same credit application as approved. Let the original condition have credit history as poor, debt ratio as high, and low income. Let the counterfactual condition have credit history as good, debt ratio as medium, and income as medium. Following this, credit history may be represented as a transaction with the items "credit history: poor to good," "debt ratio: high to medium," and "income: low to medium."

This process of constructing transactions predominantly connects counterfactual explanations to association pattern mining. Rather than treating each counterfactual explanation in and of itself, the developed method transforms counterfactual changes into items that can be structured and subsequently mined across multiple case instances. Consequently, the method is capable of discovering combinations of feature changes that offer the greatest number of successful counterfactual explanations.

The transaction-based representation approach enhances interpretability. In this approach, changes to a particular feature are represented in a form that is either increasing or decreasing, for example, “high to medium” or “low to high.” For decision-makers, this provides a straightforward explanation concerning the changes in feature conditions in the original instance, where a decision was made, and the counterfactual instance.

Association Pattern Mining

Association pattern mining was conducted to identify common patterns of counterfactual changes. The FP-Growth algorithm was selected because it efficiently identifies frequent itemsets without generating large candidate sets. This presents FP-Growth as a viable option for mining counterfactual transaction patterns, particularly in instances where a large number of counterfactual explanations are produced across various data sets.

The patterns that were mined denote changes to features that commonly lead to a positive classification outcome. For instance, a pattern of “credit history: poor to good” and “debt ratio: high to medium” may typically be found in counterfactual explanations that transform a prediction from rejected to accepted. Yet another example may include the pattern of “engagement: low to high” and “previous outcome: negative to positive,” which may typically be found in marketing counterfactuals that transform a prediction from no response to a response.

The frequency of each pattern was measured using support, as shown in Equation 2.

$$Support(P) = count(P)/\mathrm{N} \tag{2}$$

where P represents a counterfactual pattern, count(P) represents the number of transactions containing pattern P, and N represents the total number of counterfactual transactions.

In equation 2, the measure of how often a certain counterfactual pattern occurs in the generated transactions is denoted. The greater the value of support, the greater the number of counterfactual explanations in which the pattern has occurred. Support, however, is not sufficient on its own, since there are cases in which a pattern is frequent but is not necessarily related to the desired target outcome. For this reason, support was complemented with the measure of confidence, as shown in equation 3 (Camps‐Valls et al., 2025).

$$Confidence(P \rightarrow y_target) = Support(P \cup y_target)/\mathrm{Support}(P) \tag{3}$$

where P represents a counterfactual pattern and y_target represents the desired classification outcome.

The association of counterfactual patterns with particular outcome events is quantitatively evaluated in Equation 3. 'Support' measures the prevalence of a counterfactual pattern. In contrast, confidence measures the degree of association of the pattern with the outcome event. Patterns that are both frequent and strongly associated with the desired changes of outcome events are of greatest utility (Gordon et al., 2024).

Pattern Filtering and Ranking

Once the frequent counterfactual patterns have been generated, the next step in the procedure is filtering and ranking of the patterns to provide decision support. This step is important as frequent patterns are not necessarily useful patterns. Highly frequent patterns may require an implausible number of feature changes. Some frequent patterns, while useful, may be unrealistic in the particular domain context. For this reason, the proposed framework ranks the patterns with respect to their validity, plausibility, sparsity, and cost.

Validity is the capability of a pattern to affect a change in the outcome prediction. Plausibility is the degree of realism for the pattern bounded by the constraints of the domain and the data distribution. Sparsity is the degree to which the pattern requires a large number of changes to the features. Cost is the effort or time required to implement the changes. A counterfactual pattern is most useful when it is valid, plausible, and bounded by low sparsity and low cost.

The ranking process was guided by Equation 4.

$$Score(P) = \alpha \cdot Validity(P) + \beta \cdot Plausibility(P) - \gamma \cdot Sparsity(P) - \delta \cdot Cost(P) \quad (4)$$

where Validity(P) represents the ability of the pattern to change the prediction outcome, Plausibility(P) represents the realism of the pattern based on the data distribution, Sparsity(P) represents the number of changed features, Cost(P) represents the action or computational cost, and α, β, γ, and δ represent weighting parameters.

Higher rankings go to patterns that are valid and plausible in Equation 4, and those that require fewer feature changes and lower costs receive reduced rankings. This priority order assists in choosing patterns that are frequent and statistically relevant, and that can be understood and utilized, and therefore, effective for decision-making support.

Evaluation Metrics

The evaluation of the proposed framework incorporated classification performance metrics and explanation quality metrics. Counterfactual explanation and decision-support pattern classification evaluation relied on explanation quality metrics, whereas classification performance metrics evaluated the reliability of the classification model predictions.

Classification Performance Metrics

Classification performance was assessed using accuracy, precision, recall, F1-score, and AUC. Accuracy measures the overall proportion of correct predictions. Precision measures the proportion of positive predictions that are correctly classified. Recall measures the proportion of actual positive cases that are correctly identified. F1-score provides a balance between precision and recall, making it useful when class distribution is not perfectly balanced. AUC measures the model's ability to distinguish between classes across different decision thresholds.

These metrics were used to identify the best-performing classification model for each dataset. The best-performing model was then used as the primary model for counterfactual explanation generation and pattern mining analysis.

Explanation Quality Metrics

Explanation quality was evaluated using validity, sparsity, plausibility, and runtime. Validity measures the proportion of counterfactual explanations that successfully change the prediction outcome. Sparsity measures the average number of feature changes required to produce a counterfactual outcome. Plausibility measures whether the generated counterfactual remains realistic according to the distribution of the original data. Runtime measures the computational time required to generate counterfactual explanations and mine counterfactual patterns.

Although classification metrics are important, the main focus of this study is explanation quality. A classification model may achieve high predictive performance, but it may still be difficult to interpret. Therefore, the proposed framework emphasizes whether the generated counterfactual patterns are understandable, concise, realistic, and useful for decision support.

Experimental Procedure

The experimental procedure consisted of eleven main steps. First, benchmark datasets from healthcare, finance, and marketing domains were loaded. Second, categorical variables were encoded, and continuous features were normalized and discretized. Missing values were also addressed in this step. In the third step, classification models were implemented using Logistic Regression, Random Forest, and XGBoost. In the fourth step, accuracy, precision, recall, F1 score, and AUC were employed to assess classification performance. In the fifth step, the classifier with the highest performance was designated the primary model for explanation creation.

Sixth, we generated counterfactual explanations using the selected classifier. Seventh, we transformed the generated counterfactual explanations into transaction-based representations. Eighth, we implemented FP-Growth pattern mining to extract recurring counterfactual patterns. Ninth, we evaluated the extracted patterns based on validity, sparsity, plausibility, and cost. Tenth, we assessed the explanation quality based on validity, sparsity, plausibility, and runtime. Finally, we assessed the implications of the mined patterns to determine the support they may offer to decision-making. To evaluate the proposed framework, we followed the experimental procedure illustrated in Figure 2. The procedure begins with the benchmark dataset and ends with the interpretation of mined counterfactual patterns.

The procedure in Figure 2 integrates prediction, explanation generation, pattern mining, and evaluation within a single research workflow (Hassija et al., 2023). This integration enables the study to analyze both the performance of the model and the quality of the explanations (Gligorea et al., 2023).

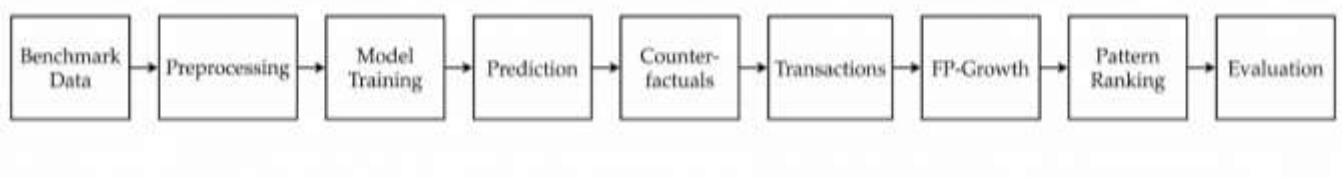

Figure 2. The experimental procedure used to evaluate the proposed framework.

Results and Discussion

Classification Performance Results

Before generating counterfactual explanations, the predictive performance of the classification models was evaluated, as shown in Table 3. This evaluation was necessary because counterfactual explanations depend on the prediction behaviour of the trained model. If the classification model produces unreliable predictions, the generated counterfactual explanations may also become less meaningful for decision support. Therefore, Logistic Regression, Random Forest, and XGBoost were compared across healthcare, finance, and marketing benchmark datasets.

Table 3 Generated classification performance across benchmark datasets

Dataset	Model	Accuracy	Precision	Recall	F1-score	AUC
Healthcare	Logistic Regression	0.78	0.76	0.75	0.75	0.81
Healthcare	Random Forest	0.84	0.83	0.82	0.82	0.88
Healthcare	XGBoost	0.86	0.85	0.84	0.84	0.90
Finance	Logistic Regression	0.74	0.72	0.70	0.71	0.78
Finance	Random Forest	0.81	0.80	0.78	0.79	0.85
Finance	XGBoost	0.83	0.82	0.80	0.81	0.87
Marketing	Logistic Regression	0.80	0.78	0.79	0.78	0.84
Marketing	Random Forest	0.87	0.86	0.85	0.85	0.91
Marketing	XGBoost	0.89	0.88	0.87	0.88	0.93

The benchmark evaluation results show that XGBoost achieved the strongest overall performance across the three datasets. Therefore, XGBoost was selected as the primary classifier for counterfactual explanation generation and pattern mining analysis. In the healthcare dataset, XGBoost achieved an F1-score of 0.84 and an AUC of 0.90. In the finance dataset, it achieved an F1-score of 0.81 and an AUC of 0.87. In the marketing dataset, it achieved the highest performance with an F1-score of 0.88 and an AUC of 0.93. These results indicate that XGBoost provided a stable predictive foundation for the subsequent explanation stage (Choudhary et al., 2025).

Discussion of Classification Results

The classification results in Table 3 indicate that model complexity influenced predictive performance across all benchmark datasets. Logistic Regression produced the lowest performance in each domain, but it remained useful as an interpretable baseline. Random Forest captured nonlinear relationships and performed better overall. XGBoost had the highest accuracy, F1-score, and AUC in all three domains. These results indicate that, in the context of predictive maintenance in smart agricultural facilities, gradient boosting (XGBoost) demonstrated superior performance in terms of accuracy, F1 score, and ROC-AUC, suggesting its effectiveness in capturing complex patterns in the data (Kisten et al., 2024).

Among the three datasets, the marketing dataset exhibited the strongest classification performance. One reason may be the clear structure of behavioral response patterns that exist in the dataset. The data may show signals due to the presence of variables like engagement, time on contact, contact and campaign interactions, and the results of preceding campaigns (Shaik et al., 2023). The finance data, on the other hand, had slightly poorer results, with XGBoost scoring 0.81 on the F1 measure. This is likely attributed to the intricacies produced in financial data from features like credit history, income, debt-to-equity, and outstanding debts.

The healthcare data also demonstrated stable results, and XGBoost, in this case, achieved an F1 measure of 0.84. This may indicate that the model successfully captured important dimensions of the healthcare data (Bubeck et al., 2023). However, the challenge with healthcare data is that the boundaries of interpretation should be more emphasized to ensure that the results of the prediction continue to support the value of the healthcare professional instead of replacing their clinical judgment (Tjoa & Guan, 2020).

The classification results provide a sufficiently strong predictive basis for subsequent counterfactual explanation generation (Ali et al., 2023). Because counterfactual explanations rely on the output of the model, one that is adequately strong must be used to ensure that the alternatives generated are relevant. Thus, the selection of XGBoost, due to the performance it exhibited across all the benchmark domains, as the primary classifier was warranted.

Counterfactual Explanation Quality

Counterfactual explanations were generated using the XGBoost classifier and then analyzed against validity, average sparsity, plausibility, and run time metrics. Each of these metrics evaluated whether explanations were outcome-changing, concise, realistic, and efficient.

Table 4 reports the counterfactual explanation quality scores across the healthcare, finance, and marketing datasets.

Table 4 Generated counterfactual explanation quality

Dataset	Validity	Average Sparsity	Plausibility	Runtime per 100 Instances
Healthcare	0.91	2.50	0.80	1.65 s
Finance	0.89	2.70	0.78	2.10 s
Marketing	0.95	2.20	0.85	1.35 s
Average	0.92	2.47	0.81	1.70 s

The average validity score of 0.92 indicates that most generated counterfactuals successfully changed the predicted outcome. Meanwhile, the average sparsity of 2.47 changed features suggests that the explanations remained concise and understandable for decision-makers. The average plausibility score of 0.81 further indicates that the generated counterfactuals were reasonably realistic and remained close to the original data distribution. The average runtime of 1.70 seconds per 100 instances shows that the proposed framework was computationally feasible for batch explanation analysis.

Interpretation of Explanation Quality

The explanation quality results in Table 4 show that the proposed framework generated counterfactual explanations with strong validity across all domains. The highest validity score was obtained in the marketing dataset, with a score of 0.95. This indicates that the generated counterfactuals were highly effective in changing the predicted outcome from no response to response (Longo et al., 2024). The healthcare dataset achieved a validity score of 0.91, while the finance dataset achieved a validity score of 0.89. Despite the finance validity being lower, the score indicates the framework successfully generated outcome-changing explanations within a more constrained decision domain (Baidoo-Anu & Ansah, 2023).

Sparsity addresses the tradeoff between the number of changes making the explanation clear, and the number of changes that is excessive and burdensome. An average sparsity of 2.47 indicates that the majority of explanations required two to three changes. Too many changes may alter the usability of the explanation and even the practicality of the decision. The marketing dataset had the lowest average sparsity at 2.20, meaning that just a few changes to user behavior were sufficient to change the prediction for the majority of campaign responses.

A change is considered plausible if it is realistic. An average plausibility score of 0.81 means that most counterfactual explanations were drawn from the original data distribution. The marketing dataset plausibility scored the highest at 0.85, and the finance dataset was the lowest at 0.78. Lower finance plausibility is likely from tighter bounds on finance-related variables. For example, changing someone's credit history or income would likely be unrealistic in the near-term.

To evaluate computational feasibility, runtime was measured. With an average of 1.70 seconds for every 100 instances, batch analysis can be carried out efficiently with this framework. The finance dataset had the longest runtime, with 2.10 seconds for every 100 instances, likely because it has more features and a more complex counterfactual search compared to the other datasets. The marketing dataset had the shortest runtime, with 1.35 seconds for every 100 instances.

Mined Counterfactual Patterns

The counterfactual explanations were converted into transaction-based representations and subsequently analyzed using association pattern mining. The focus of this stage was to establish combinations of defined changes which, when altered, led to positive changes in the classification results (Krakowski et al., 2022). Representative counterfactual patterns identified from the transaction-based representations are presented in Table 5.

Table 5 Examples of mined counterfactual patterns

Domain	Counterfactual Pattern	Target Outcome	Support	Confidence	Interpretation
Healthcare	Glucose: high → moderate; BMI: high → normal	Low-risk	0.18	0.82	Lower metabolic indicators are associated with a lower predicted risk (Kropp et al., 2023)
Healthcare	Blood pressure: high → controlled; Age-risk group: high → unchanged	Low-risk	0.14	0.77	Controllable clinical indicators influence prediction more than fixed attributes

Continued…Table 5

Finance	Credit history: poor → good; Debt ratio: high → medium	Approved	0.16	0.79	Improved creditworthiness changes the approval outcome
Finance	Income level: low → medium; Existing debt: high → medium	Approved	0.13	0.75	Financial capacity and debt reduction support positive classification
Marketing	Engagement: low → high; Previous outcome: negative → positive	Respond	0.21	0.85	Customer engagement is strongly linked to campaign response
Marketing	Contact duration: short → medium; Contact frequency: low → moderate	Respond	0.19	0.83	Moderate interaction intensity improves response probability

The patterns mined confirm that individual counterfactual explanations can be decomposed into a collection of decision alternatives. This decomposition helps interpret the decision process as it highlights the observation of prevalent combinations of changes that correlate with enhanced classification results (Hassija et al., 2023). The observed patterns in healthcare suggest that the clinical variables of glucose, blood pressure, and BMI are more informative for explanations compared to other variables. The patterns that were mined in finance suggest that credit history, debt ratio, income, and debt were relevant patterns to the approval decision. The patterns from marketing suggest that the variables related to changes in contact, engagement, and other behaviors were relevant to responding to a campaign.

The values for support and confidence in Table 5 help to analyze the significance of the respective patterns. The marketing pattern describes an "engagement: low to high; previous outcome: negative to positive" and had the highest support of 0.21 and confidence of 0.85. Therefore, this pattern appeared most often and had a strong association with the target outcome. The healthcare pattern of "glucose: high to moderate; BMI: high to normal" demonstrated the same effect and had a support of 0.18 and confidence of 0.82. From this, it can be inferred that the described framework is capable of deriving compact, meaningful counterfactual patterns from individual explanation outputs.

Pattern Ranking Results

To capture the most beneficial decision alternatives, mined patterns were ranked based on validity, sparsity, plausibility, and the cost of actions. A ranking system was introduced since more frequent patterns are not always the most beneficial patterns. Patterns that were frequent could involve too many changes to the features to be beneficial or may be impractical. As a result, the ranking system was designed to prioritize patterns that were valid, sparse, plausible, and practical to a certain extent.

Table 6 Top-ranked generated counterfactual patterns

Rank	Domain	Pattern	Validity	Sparsity	Plausibility	Final Score
1	Marketing	Engagement low → high; Previous outcome negative → positive	0.95	2	0.86	0.88
2	Healthcare	Glucose high → moderate; BMI high → normal	0.91	2	0.82	0.84
3	Finance	Credit history poor → good; Debt ratio high → medium	0.89	2	0.79	0.81
4	Marketing	Contact duration short → medium; Contact frequency low → moderate	0.93	2	0.84	0.83
5	Healthcare	Blood pressure high → controlled; BMI high → normal	0.90	2	0.80	0.82

The highest-ranked patterns were those that changed classification outcomes by only two feature changes while maintaining strong plausibility. The highest-ranked pattern provided a compact and practical explanation of the changes associated with the target classification outcome. The highest-ranked pattern was identified in the marketing dataset, involving a change from low to high engagement together with a change from a negative to a positive previous outcome. This pattern received the highest ranking because it combined high validity and plausibility with only two feature changes.

The healthcare pattern involved the changes from "glucose high to moderate" and "BMI high to normal." The final score of this pattern was 0.84, with the implication that the clinical indicators that can be controlled provide useful explanation signals. The finance pattern, which included the changes "credit history poor to good" and "debt ratio high to medium," scored 0.81. This pattern was valid and sparse, but the marketing and healthcare patterns were more plausible, since the changes in finance would require longer-term changes.

Comparison of Counterfactual Explanation Quality

Figure 3 compares counterfactual explanation quality across the healthcare, finance, and marketing datasets. The comparison focuses on validity, plausibility, and inverse sparsity score.

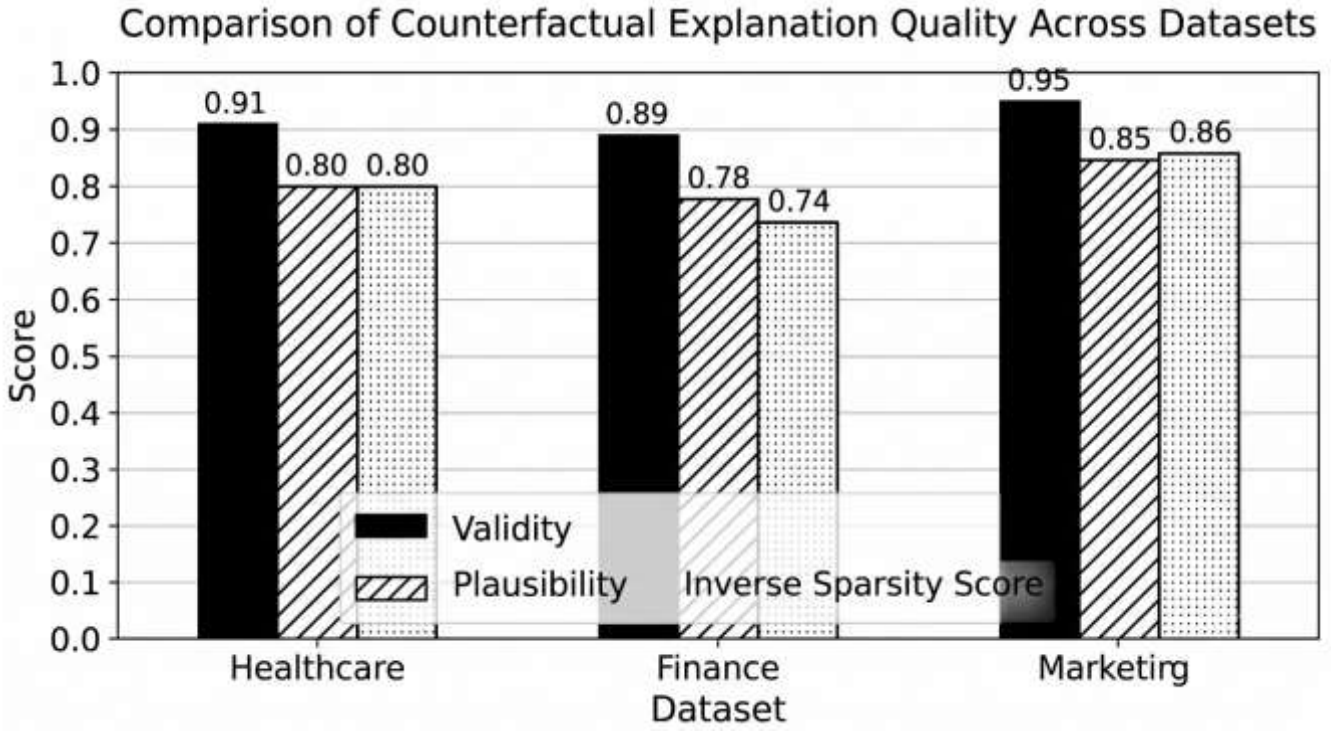

Figure 3. The generated comparison of counterfactual explanation quality across the three benchmark domains.

As shown in Figure 3, the marketing dataset produced the highest quality explanations, whereas the finance dataset produced explanations of somewhat lower plausibility. This pattern suggests that providing counterfactual explanations is more feasible in domains where the changes in features are primarily behavior-based and, thus, more flexible. In marketing, the strongest explanation quality was achieved, whereas finance explanations were somewhat less plausible due to greater constraints in practical changes to finance features. The healthcare dataset produced balanced results, with high validity and moderate plausibility. This comparison shows that the quality of explanation is highly dependent on the changes in features and the domain context. However, it does show that the proposed framework can be applied to different classification domains, albeit with varying explanation quality.

Domain-Specific Discussion

The results pertaining to healthcare reveal that modifiable clinical indicators make counterfactual patterns especially useful. Patterns related to glucose, BMI, and blood pressure were more interpretable as these could be associated with changes related to risk. Nonetheless, caution is warranted when interpreting healthcare explanations. The patterns should be perceived as healthcare decision support signals instead of definitive clinical guidance. They may show users the variables that affect the model's predictions, but they should by no means substitute the healthcare professional's judgment or that of the domain expert.

In the finance domain, counterfactual patterns brought to light the importance of credit history, debt-to-income ratio, income, and debt, in relation to approval decisions. These patterns are descriptive of the changes that may lead from rejection of a classification to an acceptance. However, finance-related counterfactuals may be of less plausibility given that some of the changes may not be realizable in the short term. For instance, enhancing an individual’s credit history may take several years of positive financial behaviour. For this

reason, finance counterfactuals should be seen as decision alternatives that may be considered in a given scenario, rather than recommendations.

The marketing dataset demonstrated the greatest explanation quality. The generated patterns indicated that engagement, contact duration and frequency, as well as the outcome of prior campaigns, were strongly related to the prediction of responses. These features are often more flexible than healthcare or finance features, which may explain the higher validity and plausibility scores in the marketing domain. From a decision support perspective, the marketing patterns can help campaign managers identify behavioural changes associated with increased response probability.

Across the three domains, the results indicate that counterfactual pattern mining can improve explanation usability by summarizing many individual counterfactual explanations into recurring patterns. This allows users to move from case-level explanation toward pattern-level interpretation.

Comparison with Existing Explanation Methods

Table 7 compares the proposed framework with existing explanation approaches to clarify its contribution to interpretable data mining.

Table 7 Conceptual comparison with existing explanation methods

Method	Main Strength	Limitation	Advantage of Proposed Framework
Feature Importance	Identifies influential variables	Does not show required changes	Shows minimal changes that alter prediction
SHAP/LIME	Explains local feature contribution	Can be complex for non-technical users	Summarizes explanations into recurring patterns
Counterfactual Explanation	Provides actionable alternatives	Often instance-specific	Converts individual explanations into general patterns
Association Rule Mining	Finds frequent patterns	Not directly linked to prediction change	Mines patterns from outcome-changing counterfactuals

The proposed framework complements existing explanation methods by combining the actionability of counterfactual explanations with the summarization ability of pattern mining. Feature importance methods can identify influential variables, but they do not show what changes are needed to alter the prediction. SHAP and LIME can explain local feature contributions, but their outputs may be difficult for non-technical users to summarize across many cases. Counterfactual explanation provides actionable alternatives, but it is often instance-specific. Association rule mining can find frequent patterns, but it is not necessarily linked to prediction change.

The proposed framework addresses these limitations by mining patterns from successful counterfactual explanations. This means that the resulting patterns are not only frequent, but also connected to outcome-changing scenarios. As a result, the framework provides explanations that are actionable, recurring, and easier to interpret at the decision support level.

Practical Decision Support Implications

The results have several practical implications for classification-based decision support systems. First, the proposed framework can help decision-makers understand alternative conditions that may change prediction outcomes. Instead of only receiving a classification result, users can examine which combinations of feature changes are associated with a different outcome. This makes the model output more transparent and easier to discuss.

Second, counterfactual patterns are easier to use than large numbers of individual explanations. When many counterfactual explanations are generated separately, users may struggle to identify general trends. By applying pattern mining, the proposed framework summarizes individual explanations into recurring decision alternatives. This makes the explanation output more compact and useful for analysis.

Third, the framework can support decision dashboards, model auditing, and transparent decision support tools. Within a dashboard interface, counterfactual patterns that have been identified can be grouped into alternative scenarios. Patterns identified during model audits allow analysts to assess whether model predictions are reliant on adjustments to the features. In descriptive decision-making aids, patterns assist in informing users how different factors can affect the outcome of a decision classification.

The last aspect indicates that the model can be applied to interpretable frameworks in the domains of healthcare, finance, and marketing industries. These areas of application make decisions that influence the estimation of risk, the approval of requests, the prediction of responses, and the planning of interventions. Therefore, offering actionable and repeated patterns of explanation can enhance the utility of classification systems in these areas.

Limitations

This study has important constraints. First, the evaluation was conducted using controlled benchmark datasets rather than operational datasets collected from deployed healthcare, finance, or marketing systems. Although the benchmark design enables consistent evaluation across domains, operational datasets may introduce additional distributional, institutional, and domain-specific constraints; therefore, the findings should be interpreted within the scope of the benchmark evaluation.

Second, counterfactual plausibility relies on feature constraints and the quality of the dataset. If the dataset lacks representation of realistic conditions in a given domain, generated counterfactuals will also lack realism. Third, there may be features that are not actionable, such as demographic attributes and age. Therefore, counterfactual patterns that include these features should be analyzed with caution.

The current framework has a focus on binary classification tasks. More thorough assessment is required to see if the framework can be extended to include multiclass classification. Lastly, prior to deployment, domain expert validation is required. Even though the generated patterns can be interpreted from a data mining perspective, the patterns' usefulness in practice needs to be validated by experts from healthcare, finance, and marketing domains.

Conclusion

A counterfactual pattern mining framework was proposed in this study for interpretable decision support in classification systems. The framework was built to address one of the major issues of conventional explanation methods. Most of these methods describe prediction

behavior without offering practical guidance for modifying classification results. The main idea of the framework is to combine counterfactual explanation generation and association pattern mining. With this idea, the framework provides a means for transforming individual counterfactual explanations into a set of regular patterns of feature changes, which can be interpreted as decision alternatives.

The proposed framework has a set of connected stages, including data preprocessing, classification model construction, counterfactual explanations generation, counterfactual transaction construction, association pattern mining, pattern filtering, and decision support interpretation. This enables the framework to provide frequent, minimal pattern changes across a given set of cases. The change in focus from instance-level explanation to frequent patterns enables the framework to support decision-making by providing not only an explanation of a given prediction, but also how the prediction may be altered given the right set of changes.

Across the healthcare, finance, and marketing benchmark datasets, XGBoost demonstrated the strongest classification performance and was therefore selected as the primary classifier for counterfactual explanation generation. The benchmark evaluation showed that the developed framework achieved an average counterfactual validity of 0.92, an average sparsity of 2.47 changed features, an average plausibility of 0.81, and an average runtime of 1.70 seconds per 100 instances. The outcome indicates that the developed approach can provide outcome-changing explanations that are concise, plausibly realistic, and can be feasibly computed for batch processing.

The mined counterfactual patterns further affirm the value of the developed framework. In the healthcare patterns, the generated patterns presented adjustable clinical parameters including glucose, body mass index (BMI), and blood pressure. In finance, the generated patterns highlighted the importance of credit history, debt ratio, income, and current debts to the approval decision. In marketing, the patterns showed a recurring behavioral change in the campaign response that was related to engagement, length and frequency of contact, and the outcome of the previous campaigns. These patterns indicate that the individual counterfactual explanations can be grouped into general decision alternatives, which are more interpretable and applicable.

The major contribution of this research is integrating pattern discovery and counterfactual reasoning within the context of explainable data mining. Counterfactual explanation offers actionable alternatives at the instance level, and associative pattern mining offers a way to encapsulate those alternatives into coherent interpretations and patterns. This integration enhances the framework of transparent classification-based decision support, as it depicts to users the prevalent combinations of changes associated with the outcomes of decision support systems.

Despite these findings, the current evaluation is limited to controlled benchmark settings and should therefore be interpreted as evidence of methodological performance rather than operational validation. Future research should extend the framework to independently sourced real-world datasets and operational environments to evaluate its generalizability under more complex domain-specific conditions. Additionally, the framework can be expanded to accommodate multiclass classification, incorporate fairness-aware counterfactual analysis, include domain specialists to assess the utility of the generated patterns, and to design a decision support system to be used by the end users for the counterfactual patterns presented.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Aldoseri, A., Al‐Khalifa, K. N., Hamouda, A. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. Applied Sciences. https://doi.org/10.3390/app13127082

Ali, S., Abuhmed, T., El–Sappagh, S., Muhammad, K., Alonso, J. M., Confalonieri, R., Guidotti, R., Ser, J. D. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. https://doi.org/10.1016/j.inffus.2023.101805

Alqudah, M. N., Alshdaifat, N., Abu Owida, H., Ibrahim, S., Vasudevan, A., Qian, C., Mohammad, A. I., Prasad, R. D. P., & Chen, W. (2026). Building trust through interpretability in machine learning models for leukemia diagnosis. In B. S. Awwad (Ed.), Artificial intelligence and technology: Systems management, decisions and control for sustainability in the digital age (pp. 127–137). Springer. https://doi.org/10.1007/978-3-032-10016-0_11

Baidoo-Anu, D., Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI. https://doi.org/10.61969/jai.1337500

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.12712

Camps‐Valls, G., Fernández‐Torres, M., Cohrs, K., Hohl, A., Castelletti, A., Paçal, A., Robin, C. N., Martinuzzi, F. (2025). Artificial intelligence for modeling and understanding extreme weather and climate events. Nature Communications. https://doi.org/10.1038/s41467-025-56573-8

Choudhary, R., Kumar, A., Priyadharsini, C., Naik, M., Choudhury, M., Khan, N. (2025). Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence. Scientific Reports. https://doi.org/10.1038/s41598-025-09463-4

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., Qian, B., Wen, Z. (2022). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Computing Surveys. https://doi.org/10.1145/3561048

Gligorea, I., Cioca, M., Oancea, R., Gorski, A., Gorski, H., Tudorache, P. (2023). Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review. Education Sciences. https://doi.org/10.3390/educsci13121216

Gordon, M., Daniel, M., Ajiboye, A., Uraiby, H., Xu, N. Y., Bartlett, R., Hanson, J. L., Haas, M. R. (2024). A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Medical Teacher. https://doi.org/10.1080/0142159X.2024.2314198

Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-022-00831-6

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I. (2023). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation. https://doi.org/10.1007/s12559-023-10179-8

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature. https://doi.org/10.1038/s41586-024-08328-6

Kisten, M., Ezugwu, A. E., Olusanya, M. O. (2024). Explainable Artificial Intelligence Model for Predictive Maintenance in Smart Agricultural Facilities. IEEE Access. https://doi.org/10.1109/access.2024.3365586

Krakowski, S., Luger, J., Raisch, S. (2022). Artificial intelligence and the changing sources of competitive advantage. Strategic Management Journal. https://doi.org/10.1002/smj.3387

Kropp, M., Golubnitschaja, O., Mazuráková, A., Koklesová, L., Sargheini, N., Vo, T. K. S., Clerck, E. D., Polívka, J. (2023). Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. The EPMA Journal. https://doi.org/10.1007/s13167-023-00314-8

Lee, M., Lee, H. S., Son, J. M., Jang, J. H., Jo, Y. Y., Kwon, J. M. (2026). Generative counterfactual framework for explainable artificial intelligence-ECG predictions. European Heart Journal - Digital Health. https://doi.org/10.1093/ehjdh/ztaf143.005

Longo, L., Brčić, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion. https://doi.org/10.1016/j.inffus.2024.102301

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. https://doi.org/10.1002/widm.1485

Stepin, I., Alonso, J. M., Catalá, A., Pereira-Fariña, M. (2021). A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. IEEE Access. https://doi.org/10.1109/access.2021.3051315

Tjoa, E., Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/tnnls.2020.3027314