Petroleum Science 22 (2025) 4157–4173

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A novel fusion of interpretable boosting algorithm and feature selection for predicting casing damage

Juan Li [a], Mandella Ali M. Fargalla [b], Wei Yan [b,*], Zi-Xu Zhang [b], Wei Zhang [b], Zi-Chen Zou [b], Tang Qing [a], Tao Yang [a], Chao-Dong Tan [b], Guang-Cong Li [b]

[a] PetroChina Dagang Oilfield Petroleum Engineering Research Institute, Tianjin, 300450, China
[b] College of Safety and Ocean Engineering, China University of Petroleum, Beijing 102249, China

ARTICLE INFO

Article history:
Received 16 July 2024
Received in revised form
5 August 2025
Accepted 6 August 2025
Available online 8 August 2025

Edited by Jia-Jia Fei

Keywords:
Casing damage
Machine learning
Feature selection
Sand production
Boosting algorithm

ABSTRACT

Casing damage resulting from sand production in unconsolidated sandstone reservoirs can significantly impact the average production of oil wells. However, the prediction task remains challenging due to the complex damage mechanism caused by sand production. This paper presents an innovative approach that combines feature selection (FS) with boosting algorithms to accurately predict casing damage in unconsolidated sandstone reservoirs. A novel TriScore FS technique is developed, combining mRMR, Random Forest, and F-test. The approach integrates three distinct feature selection approaches—TriScore, wrapper, and hybrid TriScore-wrapper and four interpretable Boosting models (AdaBoost, XGBoost, LightGBM, CatBoost). Moreover, shapley additive explanations (SHAP) was used to identify the most significant features across engineering, geological, and production features. The CatBoost model, using the Hybrid TriScore-rapper G_1G_2 FS method, showed exceptional performance in analyzing data from the Gangxi Oilfield. It achieved the highestaccuracy (95.5%) and recall rate (89.7%) compared to other tested models. Casing service time, casing wall thickness, and perforation density were selected as the top three most important features. This framework enhances predictive robustness and is an effective tool for policymakers and energy analysts, confirming its capability to deliver reliable casing damage forecasts.

1. Introduction

1.1. Background

A casing system consists of interconnected steel pipes that protect the wellbore against external forces exerted by the geological formations (Shi et al., 2021). Casing integrity is critical for the safe and efficient extraction of hydrocarbons (Deng et al., 2023). Maintaining casing integrity remains a significant challenge despite its importance due to various factors that can lead to casing damage. Practical proof from oilfield development demonstrates that the extended production period and continuous improvements to development techniques have resulted in significant casing damage in numerous oil and water wells worldwide. According to reports, the percentage of casing failures can reach up to 50%, although this information is rarely shared because of concerns about reputation, business image, and privacy. In China, wells have experienced significant damage to their casings, particularly in oilfields with clastic sandstone reservoirs that have low natural energy. The inflicted damage has resulted in the cessation of production and injection activities in impacted oil and water wells and, in certain instances, has prompted the abandonment of wells. Casing failures have a pronounced effect on oil and gas exploration, with the major oilfields in the country, including Daqing, Shengli, Jilin, and Changqing, reporting severe cases of casing damage since the 1970s. Therefore, the failure to predict and mitigate casing damage can result in significant economic losses, severe safety incidents, and harmful environmental impacts.

1.2. Literature review

1.2.1. FS for casing damage

FS techniques play an essential role in enhancing the effectiveness of predictive modeling, typically categorized into three primary groups: filter, embedded, and wrapper methods (Cai et al.,

* Corresponding author.
E-mail address: yanwei@cup.edu.cn (W. Yan).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2025.08.011

2018; Pudjihartono et al., 2022). Filter methods utilize statistical or information-theoretic criteria such as mutual information, Pearson correlation, and Euclidean distance to evaluate and rank features based solely on intrinsic data characteristics. These methods are computationally efficient but might not necessarily yield the most suitable feature subsets for specific predictive algorithms (Theng and Bhoyar, 2024). Embedded methods, by contrast, determine feature importance directly within the modeling process, as demonstrated in algorithms like Gradient Boosting (GB) and Random Forest (RF). Wrapper methods involve formulating feature selection as an optimization task closely integrated with a chosen machine learning (ML) model, thus generally achieving high predictive accuracy but potentially incurring significant computational costs. Simplified wrappers, however, can sometimes offer reduced computational demand while maintaining sufficient predictive performance (Guyon and Elisseeff, 2003). Hybrid methods merge multiple FS strategies to strike an optimal balance between computational efficiency and feature selection quality. A comprehensive understanding of these FS approaches is essential for constructing robust forecasting models. In particular, given the complexity and variability of casing damage phenomena and the dependency of predictive performance on optimal subsets of features, FS has emerged as a foundational step in numerous studies applying ML techniques within diverse energy-related contexts. Recent studies have demonstrated the growing utility of ML in casing damage prediction by integrating multidisciplinary well data and applying rigorous feature selection techniques.

For example, Zhao et al. (2020) developed a data-driven model based on a RF algorithm to predict casing damage in the Daqing Oilfield. The model was trained using diverse geological, engineering, and production parameters, preceded by extensive data preprocessing and domain-informed feature filtering. Notably, the embedded feature importance mechanism within the RF model identified key predictors, such as formation pressure, injection profile, and cumulative injection volume, as principal contributors to casing failure. Building on such efforts, Zhang et al. (2022) adopted a more structured FS framework for casing damage risk modeling in a waterflooding field. Their approach involved categorizing a broad set of candidate features into geological, engineering, and development domains, and subsequently applying two statistical filter methods, F-test and mutual information, to quantify the discriminative power of each feature. Focusing on "well-level" granularity, they systematically ranked features based on their ability to differentiate damaged from undamaged wells, thereby establishing a robust foundation for model training. In conclusion, the reviewed research consistently shows that ML models employing FS methods generally achieve better predictive outcomes than models without FS. Systematically identifying and choosing relevant features significantly enhanced these models' accuracy and generalization capability.

1.2.2. Machine learning casing damage

Research has extensively analyzed casing failures, examining the mechanisms, contributing factors, and evaluation methods for affected wells (Yin et al., 2023). In unconsolidated sandstone reservoirs, these failures are influenced by geological, production, and engineering factors. Production factors relate to fluid dynamics from formation to the wellbore, including production zones, peak daily liquid output, maximum water cut, and fluid production rate. Engineering factors concern operational data, encompassing parameters such as perforation thickness. Geological factors, derived from geologic studies and well assessments, include variables like sand layer thickness, permeability, and porosity.

Scholars have developed methods to assess and predict casing damage using geomechanics models, enhancing the understanding of the mechanical underpinnings of such damage and forecasting potential future risks (Lian et al., 2015; Lin et al., 2016; Mohamadian et al., 2021; Yang et al., 2021). These methods encompass both analytical and numerical analysis approaches. The analytical method models the casing as an ideal circle to predict damage under non-uniform external loads. In contrast, the numerical method, grounded in finite element theory, uses commercial software for spatial simulation to analyze casing deformation under complex loads and evaluate stress and deformation via numerical calculations. Despite their contributions, these methods face limitations due to vague evaluation criteria and challenges in quantitative assessment. For instance, Willson et al. (2003), Wang and Samuel (2016) utilized numerical simulations and 3D finite element models, respectively, to study casing stress under varying geological conditions. Their findings highlight the progressive stress increases on casings over time and in specific geologic scenarios. However, traditional methods, constrained by idealized assumptions and a narrow focus on specific factors, are limited in their ability to fully predict future casing damage (Mohamadian et al., 2021). Consequently, there is a growing need for new detection methods to pre-emptively identify casing damage more effectively and economically (Zhang et al., 2022).

In recent years, ML has emerged as a powerful tool for solving complex forecasting and classification problems across diverse domains, including geoscience, environmental engineering, and resource extraction (Abu-Doush et al., 2023; Braik et al., 2024; Doush et al., 2024). Within the oil and gas industry, several scholars have utilized ML methods to tackle issues related to oilfield production. For instance, Noshi et al. (2018) employed nine unsupervised algorithms, such as Bootstrap, RF, and support vector machine (SVM), to recognize the features of casing damage during drilling and fracturing. In another study, Noshi et al. (2019) utilized artificial neural network (ANN) and boosted ensemble trees to construct a prognostic model for the chance of casing failure. The framework was developed based on 26 attributes derived from drilling, fracturing, and geology data. Additionally, Song and Zhou (2019) designated ten significant parameters that affect casing damage, such as a sand layer, casing, and perforation information. They established a model for casing damage risk calculation using Gradient Boosting Decision Tree (GBDT), and their model achieved a prediction accuracy of 86.3%. In their study, Tang et al. (2019) identified 19 influential factors related to casing damage and developed a hazard prediction model employing Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) algorithms. They used 23 distinct features to predict casing damage occurrences in the Gangxi Oilfield. Among these features, perforation density and manufacturing pressure differential emerged as critical determinants influencing casing integrity. Comparative analysis revealed that XGBoost outperformed LightGBM, achieving 99% and 94% prediction accuracies, respectively. Li et al. (2024) developed a multi-factor casing damage prediction method based on six ML models. Overall, these studies demonstrate the effectiveness of data-driven methods for addressing challenges in oilfield production.

Despite recent advancements, a critical gap remains in the prevailing literature: a perceptible scarcity of sufficiently interpretable ML models for forecasting casing damage, an essential requirement for enabling transparent, data-driven decision-making in well integrity management. Many existing studies overlook feature selection entirely or adopt naïve, single-method approaches that may retain redundant or irrelevant variables, thereby reducing model clarity and generalizability.

To address these limitations, this study proposes a comprehensive and interpretable ML framework for casing damage prediction. The approach incorporates a multi-stage feature selection

strategy, combining a novel TriScore filter method, a wrapper-based selection, and their hybrid integration, to identify the most relevant predictors while minimizing redundancy systematically. This is complemented by expert-driven refinement and rigorous model development using multiple boosting algorithms with hyperparameter optimization via Bayesian search. Model performance is evaluated on training and test datasets using standard classification metrics, and further enhanced through shapley additive explanations (SHAP) analysis to ensure transparency and interpretability. This framework addresses the shortcomings of prior studies that rely on naïve or single-method feature selection and ensures the development of robust, high-performing, and explainable predictive models tailored to support informed well integrity decision-making.

1.3. Novelty and contributions

Our study comprehensively contributes to ML-based casing damage forecasting, encompassing multiple aspects. Our primary aims are to boost black-box ML models' interpretability and predictive precision.

(1) This study innovatively utilizes high-level casing damage data, a complex task due to the generally limited dataset availability across complete oil and gas sources, particularly over casing damage. On the other hand, other researchers (e.g., Mohamadian et al., 2021; Wang et al., 2023; Xue, 2020) have primarily utilized low-level data with fewer features, casting doubts on the reliability of their conclusions.
(2) This framework enhances the accuracy of our ML-driven predictions and offers novel insights into their features. Our study examines the use of many sources of casing data in China, performing a detailed analysis of various aspects such as engineering, geological, and production parameters. This represents a notable deviation from past studies, which usually depend on an inadequate group of pre-determined features short of sufficient rationale.
(3) This research distinguished itself from this study (Li et al., 2024) by performing a comprehensive and comparative FS investigation. This unique framework integrates three TriScore FS techniques, as described in this research, with a wrapper method incorporating four Boosting ML models. This enables a comprehensive comparison study.

The structure of the paper is as follows: Section 2 presents the methodology employed in this study. Section 3 discusses the key factors influencing casing damage and details the data preprocessing steps undertaken prior to model training. Section 4 presents the results and discussion, compares the performance of the proposed model with other ML approaches, summarizes key findings, and outlines policy implications and study limitations. Finally, Section 5 concludes the research and suggests directions for future work.

2. Proposed approach

Fig. 1 presents a comprehensive workflow for predicting casing damage using supervised ML, integrating domain expertise, feature selection techniques, and performance evaluation mechanisms. The process begins with a literature review to identify recent advancements in ML-based casing damage prediction, followed by the development of a comprehensive list of candidate features derived from engineering, geological, and production datasets. After data collection, the dataset undergoes preprocessing to address missing values, outliers, and inconsistencies. The next stage involves applying three distinct FS approaches: (1) TriScore FS (a hybrid filter-based method incorporating minimum redundancy maximum relevance (mRMR), RF, and F-test), (2) wrapper methods, and (3) a hybrid TriScore-wrapper strategy.

These methods reduce the feature space while retaining the most relevant predictors. Expert input is also incorporated at this stage, allowing for manual inclusion of domain-relevant features that may have been filtered out algorithmically. The selected features are then used to train various boosting-based ML models (e.g., XGBoost, LightGBM, CatBoost), and Bayesian optimization is employed to fine-tune model hyperparameters. Model performance is evaluated on a training set, and the best-performing models are identified based on metrics such as accuracy, precision, recall, and F1-score, followed by validation on a separate test set. Finally, to enhance model interpretability, SHAP analysis is applied to extract and visualize the importance of the selected features, providing insights into their contribution to casing damage prediction. This end-to-end framework ensures a balance between model accuracy, robustness, and interpretability, which are key requirements for informed decision-making in well integrity management. This leads to the creation of a feature subset that optimally balances the accuracy of the ML model with its interpretability, as detailed by Chen et al. (2023).

The architecture of our proposed method comprises several crucial steps. Fig. 2 illustrates the proposed hybrid feature selection framework used for developing interpretable and high-performing ML models for casing damage prediction. The process begins with the complete set of features denoted as G_0, which includes all available geological, engineering, and operational variables. This initial feature pool undergoes a two-stage selection process. First, the TriScore FS module employs a filter-based approach to evaluate and reduce the original feature space, resulting in a subset G_1 of statistically relevant features. A wrapper FS method refines this reduced set, which evaluates feature subsets based on their performance within specific ML models. This step yields an optimized feature group G_2, balancing statistical relevance and predictive contribution. The final selected feature subsets (G_1 and G_2) are then input into a suite of supervised ML models, including multiple boosting algorithms (e.g., Adaptive Boosting (AdaBoost), XGBoost, LightGBM, CatBoost), to assess predictive performance using training data. This framework aims to ensure the resulting models are accurate and interpretable by identifying the most informative and non-redundant features through an integrated filter-wrapper selection pipeline.

2.1. Cross-correlation analysis

This study examined the relationships between affecting factors and each prediction target using 3 analytical opinions: PCC, SCC, and KCC. The PCC (Dai et al., 2024; Fargalla et al., 2024; Liu et al., 2024) correctly measures the degree of linear relationship between two variables, X and Y. The equation provided below serves as the definition:

$$P = \frac{A(XY) - A(X)A\backslash(Y)}{\sqrt{A\left(X^2\right) - A^2(X)}\sqrt{A\left(Y^2\right) - A^2(Y)}} \quad (1)$$

In this expression, $A(X)$ and $A(Y)$ denote the mean values of the variables X and Y, respectively. $A(XY)$ represents the mean of the product of corresponding values of X and Y, while $A(X^2)$ and $A(Y^2)$ represent the mean of the squared values of X and Y, respectively. The terms $A^2(X)$ and $A^2(Y)$ refer to the square of the mean values of X and Y.

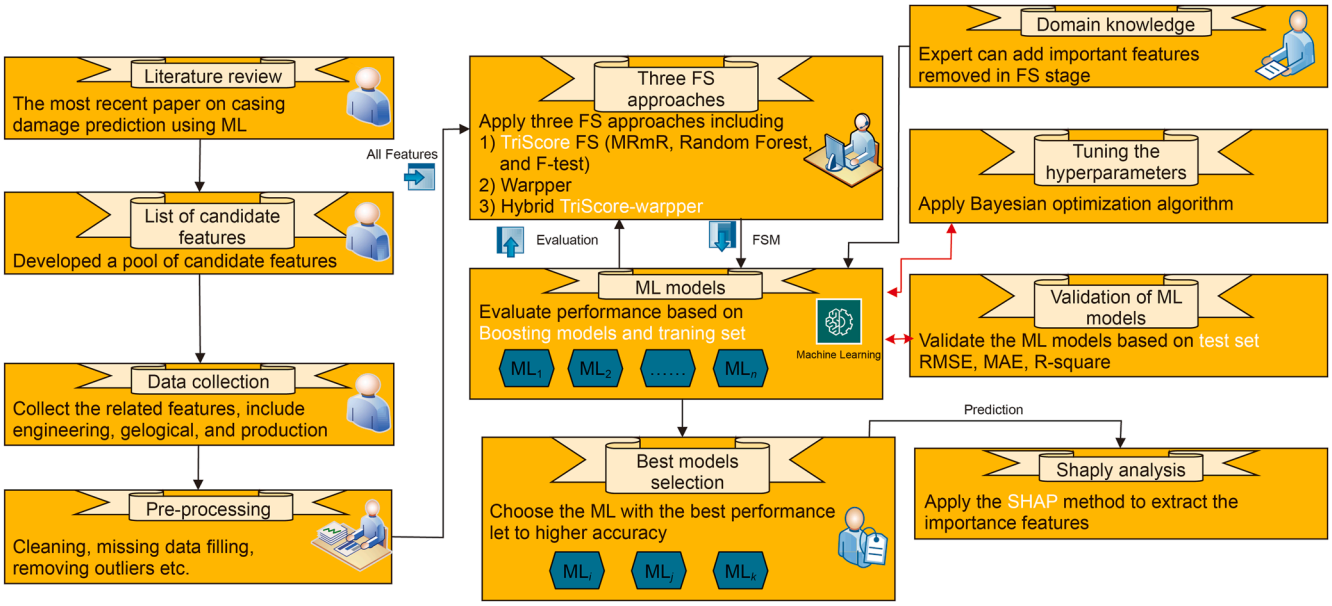

Fig. 1. The proposed framework combines two FS techniques and interpretable ML models to predict casing damage.

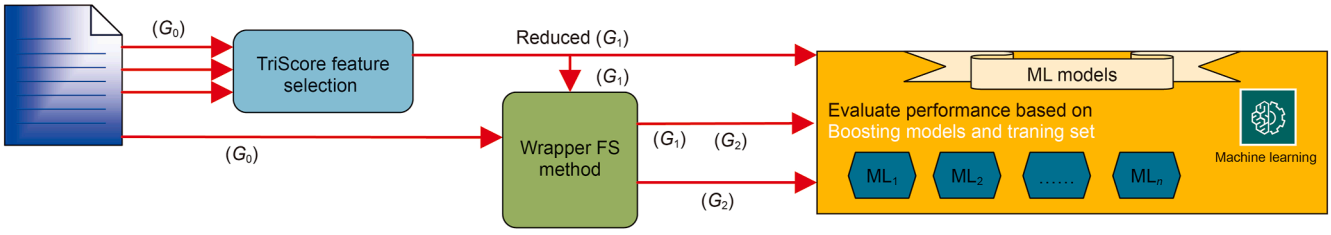

Fig. 2. The methodology of FS employs three distinct techniques.

The SCC (Wang et al., 2024) evaluates the monotonic association between variables by analyzing the linear relationship by ranking the two factors. The equation for computing the SCC for an individual data set with a sample size of n is given by:

$$\beta = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

Here, d_i is the rank variance among X_i and Y_i.

The Kendall correlation coefficient denoted by σ, is a rank relationship metric utilized to assess the overall relationship among factors, encompassing both monotonic and non-monotonic relationships. The formula for the Kendall correlation coefficient is provided below:

$$\sigma = \frac{C-D}{\frac{1}{2}n(n-1)} \tag{3}$$

In this context, 'C' represents the number of couples of factors in the instance data from X and Y that are concordant (two factors forming a consistent pair). At the same time, 'D' stands for the quantity of discordant pairs (elements in a pair that are inconsistent). Each of the three correlation coefficients discussed analyzes the correlations from distinct perspectives, providing a comprehensive overview of the relationships in the input data.

2.2. Feature selection methods

2.2.1. F-test

FS method ranks input features according to their implication for a precise ML function by utilizing statistical investigation to evaluate the importance and impact of each feature on the output. The significance of features is quantified using the F-statistic (F). The F-value quantifies the significance of a characteristic in accounting for the inconsistency of the yield feature. The process of FS using the F-test consists of three phases: determining F-values for individual features, organizing these factors in a downward order based on F-values, and choosing the top-k features with the maximum F-values. The user can either specify the value of k or select it over cross-validation. The F-test FS method identifies and eliminates redundant or less significant features, resulting in an added streamlined and precise ML model. This method is especially beneficial for handling datasets with many dimensions. FS can improve the model's performance and reduce the risk of overfitting.

2.2.2. mRMR

Maximum relevance and minimum redundancy pertain to minimizing unnecessary repetition or duplication in data. The mRMR technique is employed to select features that optimize the relevance of input characteristics concerning the output feature. Concurrently, this method eliminates redundant inputs, thereby enhancing the efficiency and effectiveness of feature selection (Zhang et al., 2023). The method employs mutual information (MI) to evaluate the significance and duplication of characteristics. The MI is demarcated in the following manner:

$$I(A/B) = \iint p(a,b)\log\frac{p(a,b)}{p(a)s(b)} \tag{4}$$

In this formula, A and B denote vectors, $p(a,b)$ represents the shared probability density, and $p(a)$ and $p(b)$ are the marginal

probability densities. Given a feature group G with m features (x_i, $i \in (1, m)$), the max relevance criterion seeks a subgroup with the maximum importance to the output feature y, as illustrated below:

$$\max D(G, y), D = \frac{1}{|G|} \sum_{x_i \in G} I(x_i, y) \tag{5}$$

To detect irrelevant features, the minimum redundancy criterion evaluates potential redundancies within the max relevance selected features, as demonstrated below:

$$\max R(S), R = \frac{1}{|S|} \sum_{x_i \in S} I\left(x_i, x_j\right) \tag{6}$$

An iterative search algorithm finds the best solution that meets both restrictions. Given an existing feature set S_{m-1}, the objective is to select the mth feature from the set $\frac{X}{S_{m-1}}$, as described in Eq. (4).

$$\max_{x_{j \in X - S_{m-1}}} \left[I\left(x_j, y\right) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I\left(x_i, x_j\right) \right] \tag{7}$$

2.2.3. Random Forest

RF is an incredibly efficient ensemble learning method utilized in regression and classification assignments for FS. The training procedure incorporates a collection of multiple Decision Trees (DTs), with each DT making an independent contribution to the feature selection. This independence helps reduce the overfitting typically seen in single DT. Random Forest is particularly advantageous for datasets with varied characteristics. In the Random Forest framework, FS is achieved by distributing data subsets across individual trees. The RF algorithm assesses the importance of features and aids in choosing a feature by using fundamental ideas from the DT approach, including variance Eq. (8) and knowledge gain Eq. (9).

$$\text{Entopy}(p_1, p_2, \ldots, p_k) = -\sum p_1 \log_2(p_1) \tag{8}$$

$$I_G = 1 - \sum_{J=1}^{C} p_J^2 \tag{9}$$

2.2.4. TriScore feature selection (TriSFS) proposed in this paper

One of the main difficulties in FS is selecting the best proper filter and integrated feature selection FS algorithms for a particular dataset. Due to the varied logic and statistical measures underlying different FS techniques, they often select distinct feature sets. A feature deemed significant by one method may be less important in another. However, it is essential to note the inherent limitations of these methods: filter methods may fail to recognize feature interdependencies, and embedded methods depend heavily on ML models. TriSFS is utilized to improve FS accuracy and robustness. This technique integrates results from multiple FS methods to better determine feature retention or exclusion, leveraging their strengths and compensating for their weaknesses. The process of the proposed ensemble feature selection technique is illustrated in Fig. 3. The ensemble feature selection process involves.

(1) Parameter setting: initiate with N potential features and choose a subset size (m) depending on the number of votes (v) received from K basic FS techniques.
(2) Base FS method selection: select a combination of filter and embedded feature selection algorithms.
(3) Feature ranking and scoring: calculate feature importance scores and iteratively eliminate the least important until reaching the desired feature count.
(4) Importance aggregation: integrate scores from various feature selection methods through weighted averages. Alternatively, a voting threshold is that the feature must appear in at least two methods.
(5) Final selection: keep the top 'm' features as determined by the highest scores and requisite votes, discarding others.

2.3. Wrapper feature selection

The SIFE method, renowned for its effectiveness, is explicitly considered for FS in high- and low-dimensional spaces (Karasu et al., 2020). SIFE enhances its search efficacy through a novel triparental recombination technique based on set concept operations such as 'union' and 'cross-section'. It integrates fuzzy granulation to facilitate population initialization and elite selection. This integration fosters intergenerational variety and decreases the necessity for comprehensive suitability assessments. The basic objective of SIFE is to achieve an ideal equilibrium between discovering novel results and exploiting established ones while ensuring that the computational difficulty remains reasonable. SIFE's effectiveness in navigating and optimizing diverse search spaces underpins its selection for this study.

SIFE employs a methodical technique in every iteration to assess the ranks of solutions for selection and repetition. One of the significant aims of feature selection is to optimize a quality metric that fulfills two fundamental aims of ML algorithms: minimalizing the algorithm's fault measured and selecting a concise subset of significantly related and less duplicate features. In order to accomplish this goal, the SIFE basis provides a purposeful formula:

$$\text{minimize } \mathrm{F}(d_i) = w_1 \times Er(d_i) + w_2 \times Ld(d_i) \forall d_i \in \Omega, \tag{10}$$

This function assesses the factor subset d_i in the range of possible factors Ω, since the metrical Er, and the ratio of certain factors Ld. The weights w_1 and w_2 are assigned standards of 0.80 and 0.20, respectively, reflecting their relative rank, given the study's focus on a limited number of features.

2.4. ML models

This section examines four notable Boosting ML models: CatBoost, XGBoost, AdaBoost, and LightGBM. A thorough literature review has established these diverse ML models as effective for classifying casing damage.

2.4.1. Adaptive Boosting (AdaBoost)

AdaBoost method enhances model performance by focusing on areas where initial iterations underperform. AdaBoost performs an iterative process to boost the performance of weak classifiers and turn them into robust classifiers. It achieves this by applying a Bayesian classifier strategy to effectively decrease the chances of misclassification. This is done by combining many weak classifiers (Wang et al., 2018). The procedure begins by constructing an initial classifier from an unweighted training sample, such as a DT. Each subsequent iteration adjusts to highlight and correct potential misclassifications, increasing the weight of misclassified instances to ensure they are addressed in the next cycle. This method iteratively combines several weak learners, adjusting the training focus based on prior errors, to form a robust classifier that effectively distinguishes between classes.

2.4.2. Extreme Gradient Boosting (XGBoost)

XGBoost represents a sophisticated version of the optimized Gradient Boosting algorithm, which is noted for its high efficiency,

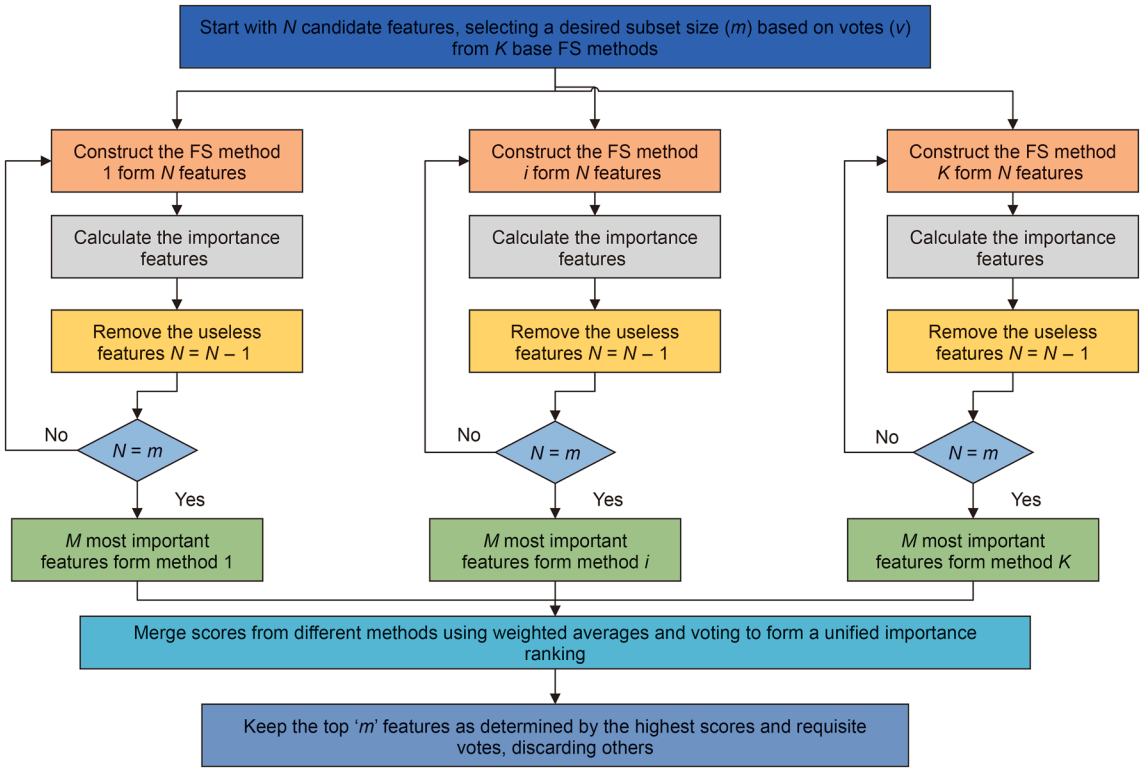

Fig. 3. The concept of the proposed TriSFS.

flexibility, and portability. As a tree-based, supervised ML algorithm, XGBoost is applicable to classification and regression tasks, with a primary focus on its use for classification. XGBoost enhances the conventional Gradient Boosting Machine (GBM) framework through several system optimizations and algorithmic enhancements: (1) It employs a parallelized tree-building process, constructing trees sequentially while utilizing parallel computations. (2) It features a tree pruning technique where trees are grown to maximum depth and then pruned back based on a loss function threshold. (3) XGBoost utilizes cache awareness and out-of-core computing techniques to handle computational time and memory capacity effectively. (4) Regularization techniques are integrated to prevent overfitting, regulating the model by constraining or reducing coefficients towards zero (Li et al., 2024; Ren et al., 2023). (5) It efficiently handles missing values, and (6) features an inherent cross-validation mechanism, which obviates the need for external cross-validation steps and helps specify the required number of iterations directly. Despite its advantages, XGBoost requires extensive parameter tuning due to its high flexibility. It incorporates cutting-edge regularization (L1 & L2) to enhance model generalization. Furthermore, XGBoost offers superior performance relative to traditional Gradient Boosting, with its training processes notably faster and capable of parallelization across clusters.

2.4.3. Light Gradient Boosting Machine (LightGBM)

The LightGBM method integrates two novel techniques, Gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB), to handle extensive datasets and high-dimensional feature spaces. GOSS prioritizes instances with significant gradients while randomly selecting instances with lesser gradients. EFB merges multiple exclusive features into fewer, reducing unnecessary computations for features with zero values. LightGBM discretizes continuous features as a histogram-based algorithm, enhancing training speed and efficiency and reducing memory usage. Contrary to the traditional depth-wise expansion of decision trees, LightGBM grows tree leaf-wise (best-first), opting to split leaves that can significantly reduce losses, thus potentially yielding lower losses compared to level-wise tree growth. Although more prone to overfitting, the leaf-wise approach is advantageous for larger datasets due to its flexibility.

2.4.4. Categorical boosting (CatBoost)

CatBoost, a combination of the words 'Category' and 'Boosting', is specifically developed to handle data that includes category, numeric, and text elements. It demonstrates exceptional proficiency in handling categorical data and datasets of limited size. CatBoost utilizes a symmetric or oblivious tree structure where every tree level applies the same characteristics to divide the training sample into right and left partitions. This results in a tree with a depth of k and precisely 2^k leaves. The technique builds decision trees sequentially, where each tree is designed to minimize the loss compared to the previous one. The initial parameters control the number of trees to help reduce overfitting. Based on the specified training settings, CatBoost can also stop training early if overfitting is detected (Zhou et al., 2024).

2.5. Bayesian optimization algorithm

The Bayesian optimization algorithm is recognized as a broad optimization technique specifically designed to manage costly objective functions. It sets itself apart from conventional approaches by functioning autonomously without relying on population-based and genetic operators like selection, mutation, and crossover. This approach uses a Gaussian method to estimate an acquiring function,

accurately forecasting the performance of the goal function (Awal et al., 2021). Moreover, Bayesian optimization improves its efficacy over time by incorporating accumulated historical data and using previously gathered statistics to refine its search for optimal solutions. According to the literature, Bayesian optimization is more effective than both grid search and random search, and it competes favorably with modern evolutionary optimization algorithms.

2.6. ML analysis by SHAP

The lack of interpretability in black-box ML models has increased criticism, emphasizing the necessity for quantitative examination of the correlation between input and output features in decision-making processes. SHAP provides a thorough approach to evaluating these models by utilizing the standard shapley value in game theory to connect optimum credit distribution to local reasons. The model's predictions are delineated by aggregating the impacts of different variables, boosting comprehension of the relevance of each element and facilitating effective decision-making. SHAP values are calculated using a linear clarification model as an explainable proxy for the ML model.

$$g(z') = \gamma_0 + \sum_{j=1}^{M} \gamma_j z'_j \tag{11}$$

Let g represent the explanation model. This point $z' \in \{0,1\}^M$ indicates the coalition vector, M is the maximum coalition size, and γ_j denotes feature attribution for a feature j-th. SHAP values assess feature rank by associating the forecast of model performance with and without separate features across different feature combinations, as illustrated in Eq. (12):

$$\gamma_0 = \sum_{G \subseteq z' \{i\}} \frac{|G|!(M-|G|-1)}{M!} [f_x(G \cup \{i\}) - f_x(G)] \tag{12}$$

In this formula, G represents the group of features for which z' is not equal to zero, and $f_x(G) = E[|f(x)x_G|]$ denotes the predicted model output of $f(x)$ when impacted by the features in G.

3. Data preparation and preprocessing

3.1. Data description

The data utilized in this work comprises the geological parameters, engineering parameters, production data, and casing damage information of 244 production layers in 133 wells located in the Gangxi Oilfield. A total of 68 production layers in 64 wells had casing damage. The rate of casing damage for the production layers was 27.9%. The data required in Table 1 are obtained by extracting information from the current database of the Gangxi Oilfield to create a sample set of casing damage. The sample set includes text kinds for casing steel grade and oil reservoir group, with each type represented by a distinct number. The numbers 1, 2, 3, and 4 correspond to the casings of steel grades J55, K55, N80, and P110, respectively. The numbers 1 and 2 denote the Ming II and Ming III reservoir groupings. Furthermore, any missing data were fully incorporated, and any inaccurate data were corrected. The data statistics overview of these 244 production tiers is presented in Table 1. Fig. 4 illustrates the statistical distribution of the standardized dataset. It displays a blue numeric data distribution diagram and an orange textual data distribution diagram.

3.2. Data preprocessing

Although wells with relatively complete data were carefully selected, missing data issues persisted, as illustrated in Fig. 5 and Table 2. Therefore, additional data processing was necessary to improve data quality. In this study, missing values and outliers were managed using a combination of field expertise and standard ML techniques. Minor missing values were addressed through outlier correction methods, whereas extensive gaps were manually filled based on practical experience and oilfield-specific knowledge. Despite being time-intensive, these steps were crucial, as high-quality data directly underpins the accuracy and reliability of predictive models. Data normalization is an essential process that involves mapping data onto a unit sphere and dealing with variances in the scales of feature dimensions. The normalization process, denoted by $x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, standardizes data points x within the range of minimum ($x_{\min}$) and maximum ($x_{\max}$) values to improve the performance of ML algorithms that rely on Euclidean distance metrics.

3.3. Model evaluation

After deploying a ML algorithm, it is crucial to assess its performance using specific evaluation metrics. In this study, the metrics employed for classification are accuracy and recall. Accuracy

Table 1
Statistics of influencing factors of casing damage in production wells.

No.	Abb.	Feature name	min_vals	max_vals	Range	Variance	Std. dev
1	F1	Perforation top	827	1408.7	581.7	11177.8	105.7
2	F2	Perforation bottom	833.8	1412.4	578.6	11178.2	105.7
3	F3	Perforation thickness	1	14	13	4.8	2.2
4	F4	Perforation density	10	32	22	18.3	4.3
5	F5	Perforator phasing	90	135	45	180.3	13.4
6	F6	Casing wall thickness	6.2	9.17	2.97	0.6	0.8
7	F7	Casing steel grade	–	–	–	–	–
8	F8	Sand layer top	826.9	1408.7	581.8	11204	105.8
9	F9	Sand layer bottom	833.8	1412.4	578.6	11150.5	105.6
10	F10	Casing service time	0.48	50.34	49.86	141.3	11.9
11	F11	Sand layer thickness	1	19.2	18.2	10.1	3.2
12	F12	Permeability	162.9	5431.1	5268.2	561182.6	749.1
13	F13	Porosity	20.2	72.09	51.89	21.88086109	4.677698268
14	F14	Maximum daily liquid production of the single-layer	1.21	359.7	358.49	1158.6	34
15	F15	Maximum water cut	28.7	100	71.3	163	12.8
16	F16	Maximum fluid production intensity	0.311111111	67.36	67	84	9.2
17	F17	Qlcum	19.59	2123576	2123556.41	26882394301	163958.5
18	F18	Qlmax/H	7.99	7391.9	7383.9	294620.3	542.8

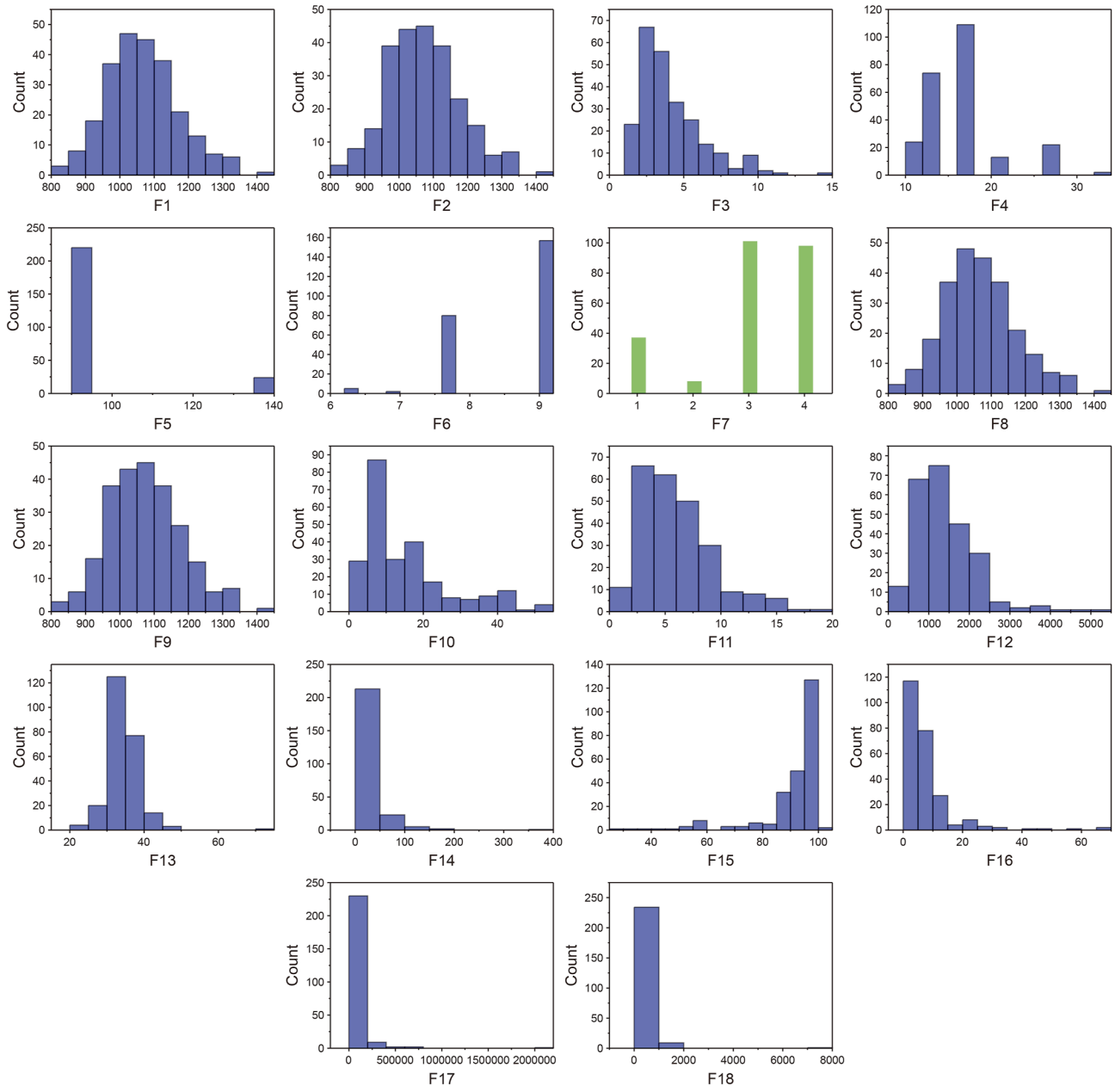

Fig. 4. Statistical distributions of all features.

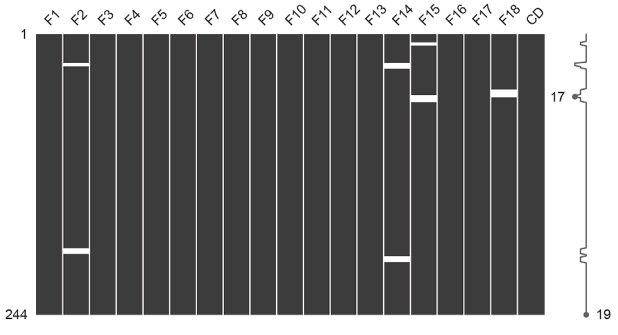

Fig. 5. Visualization of the missing data.

measures the overall correctness of the model, while recall emphasizes the model's ability to correctly identify positive cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

where, TP (true positive) is the number of correctly predicted positive cases. TN (true negative) is the number of correctly predicted negative cases. FP (false positive) is the number of negative cases incorrectly predicted as positive. FN (false negative) is the number of positive cases incorrectly predicted as negative.

Table 2
Summary of the missing values.

Model feature	Missing count	Missing, %	Non-missing count	Non-missing, %
F1	0	0	244	100
F2	8	3	236	96.7
F3	0	0	244	100
F4	0	0	244	100
F5	0	0	244	100
F6	0	0	244	100
F7	0	0	244	100
F8	0	0	244	100
F9	0	0	244	100
F10	0	0	244	100
F11	0	0	244	100
F12	0	0	244	100
F13	0	0	244	100
F14	10	4	234	95.9
F15	9	3.6	235	96
F16	0	0	244	100
F17	0	0	244	100
F18	7	2	237	97
CD	0	0	244	100

4. Results and discussion

The experiments were conducted using a computer equipped with an Intel(R) Core i7-12700u processor, 16.0 GB RAM, and NVIDIA GeForce MX150 GPU, utilizing Python 3.9.13. The included packages are NumPy 1.21.5, Pandas 1.4.4, Scikit-Learn 1.0.2, and TensorFlow 2.11.0.

4.1. Hyperparameter setting

Hyperparameter optimization is an essential aspect of ML since it is crucial for maximizing the possibility of forecast models and achieving optimum performance. The systematic and sensible exploration of the most appropriate hyperparameter scenarios is vital in ML research. For this paper, we allocated 80% of the dataset for training and adjusting the hyperparameters, while the residual 20% was set aside as the testing set. The process of tuning hyperparameters was carried out by utilizing Bayesian optimization, which was implemented using the statistics and ML toolbox of Python 3.9.13. The findings obtained from the hyperparameter optimization technique using Bayesian optimization are presented in Table 3. Overfitting is the phenomenon that occurs when the model fits the training set excessively well. The cross-validation technique prevents the ML model from being trapped in a suboptimal solution and mitigates the risk of overfitting. The study employed *K*-fold cross-validation. The value of *K* in this study is 5.

4.2. Cross-correlation analysis results

The findings of the cross-correlation analysis are displayed in Figs. 6–8. Upon analyzing the data, utilizing three primary correlation coefficients— linear, monotonic, and general correlations, it became evident that the relationships between predicted targets and impacting variables remained constant, despite being assessed using distinct correlation coefficients.

Using Pearson (PCC), Spearman (SCC), and Kendall (KCC) correlation matrices, we identified high correlations among several feature pairs, notably (F1, F2) and (F8, F9), which consistently exceeded a threshold of 0.90 (indicating “strong” or “high” correlation). Despite these findings, an extended evaluation (see Appendix A) confirmed that removing or retaining such correlated features did not materially affect model accuracy or predictive metrics. This outcome aligns with prior research showing that tree-based ensemble methods (e.g., AdaBoost, XGBoost, LightGBM, CatBoost) are relatively robust to multicollinearity, allowing us to preserve these features based on their domain relevance.

4.3. Feature selection results

This section analyzes the results of the feature selection algorithms, emphasizing the 12 essential aspects that should be included in our model. Table 4 records the specific results of different FS algorithms. This table displays the major characteristics used in the study after eliminating any collinearity. It includes columns for mRMR, RF, and F-test, which indicate the relative relevance of features assessed by each feature selection method.

As an illustration, in the mRMR algorithm, the variable “casing service time” (F10) has a relative importance of 0.281, suggesting its utmost significance. The ‘maximum water cut’ (F15) follows with a significance rating of 0.107, ranking it second. The “vote” row emphasizes each feature’s relevance, indicating its cumulative importance across the FS methods: RF, mRMR, and F-test. Furthermore, the ‘mean importance’ row calculates the average relevance score of features across different methodologies. For example, the variable “casing wall thickness” (F6) has a relevance score of 0.024 in mRMR and much higher values in RF and F-test, at 0.237 and 1, respectively. By employing this improved process, the ultimate assessment of feature significance, determined by calculating average scores, yields a hierarchical arrangement of the characteristics. The attributes “casing wall thickness” (F6), “casing service time” (F10), and “casing steel grade” (F7) have been determined to be the most important, ranking first, second, and third accordingly, according to the aggregate rankings obtained from the mRMR, RF, and F-test methodologies. The difference in the importance of features emphasizes the crucial function of ensemble feature selection strategies that utilize different filters and integrated feature selection methods. Recognizing this variation, we aim to use the advantages of several approaches while mitigating the potential of choosing an unfavourable combination of characteristics. To enhance our research design, we have expressly incorporated the 12 most significant features recognized by the TriScore technique into the input feature subgroup for the

Table 3
Boosting the hyperparameter setting.

Algorithm	Parameter tuning	Optimum parameter	Advantage
AdaBoost	n_estimators = {100–900} learning_rate = [0.001,0.01,0.1]	n_estimators = 400 learning_rate = 0.1	Adaptively adjusts and tries to self-correct in each iteration of the boosting process.
XGBoost	n_estimators = {100–900} learning_rate = [0.001,0.01,0.1] gamma = 0 max_depth = {2,6}	n_estimators = 400 learning_rate = 0.1 gamma = 0.6 max_depth = 4	Attractive for big and small data applications
LightGBM	n_estimators = {100–900} learning_rate = [0.001,0.01,0.1] max_depth = {2,6}	n_estimators = 400 learning_rate = 0.002 max_depth = 4	Sensitive to overfitting
CatBoost	n_estimators = {100–900} learning_rate = [0.001,0.01,0.1]	n_estimators = 400 learning_rate = 0.1	Handle missing values and imbalanced data internally.

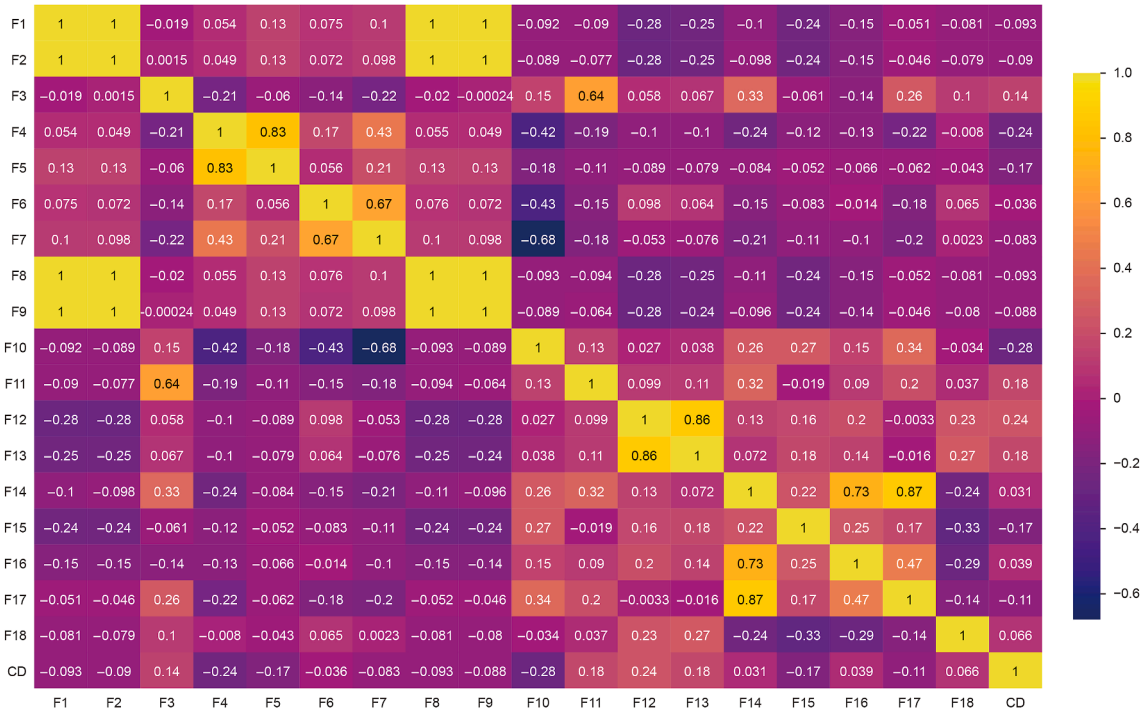

Fig. 6. Pearson correlation coefficient.

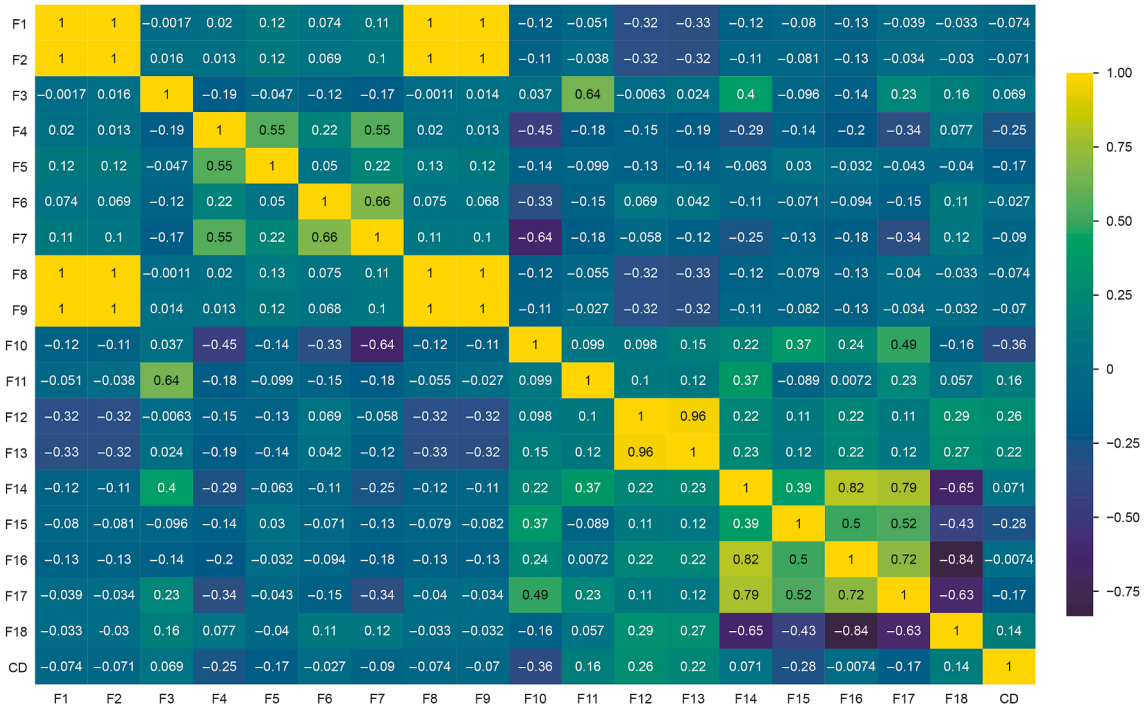

Fig. 7. Spearman correlation coefficient.

TriScore-wrapper approach. This modification aligns our approach with well-recognized best practices in feature selection.

The feature selection outcomes for the ML models designed to predict casing damage are detailed in Table 6. The table uses the sign "√" to indicate the insertion and "×" to mark the omission of specific features. A consensus was established after identifying the 12 most crucial features (n) using a TriScore of F-test, mRMR, and RF methods (as shown in Table 5). The efficacy of each of the four ML models is assessed using a set of 12 important features derived from the TriScore method's selected feature subset (G_1). In contrast, the wrapper method (G_2) often selects a smaller subset from the 12 features yet displays a broader variety of features across different ML algorithms. The voting threshold is that the feature must appear in at least three methods. Notably, the ML

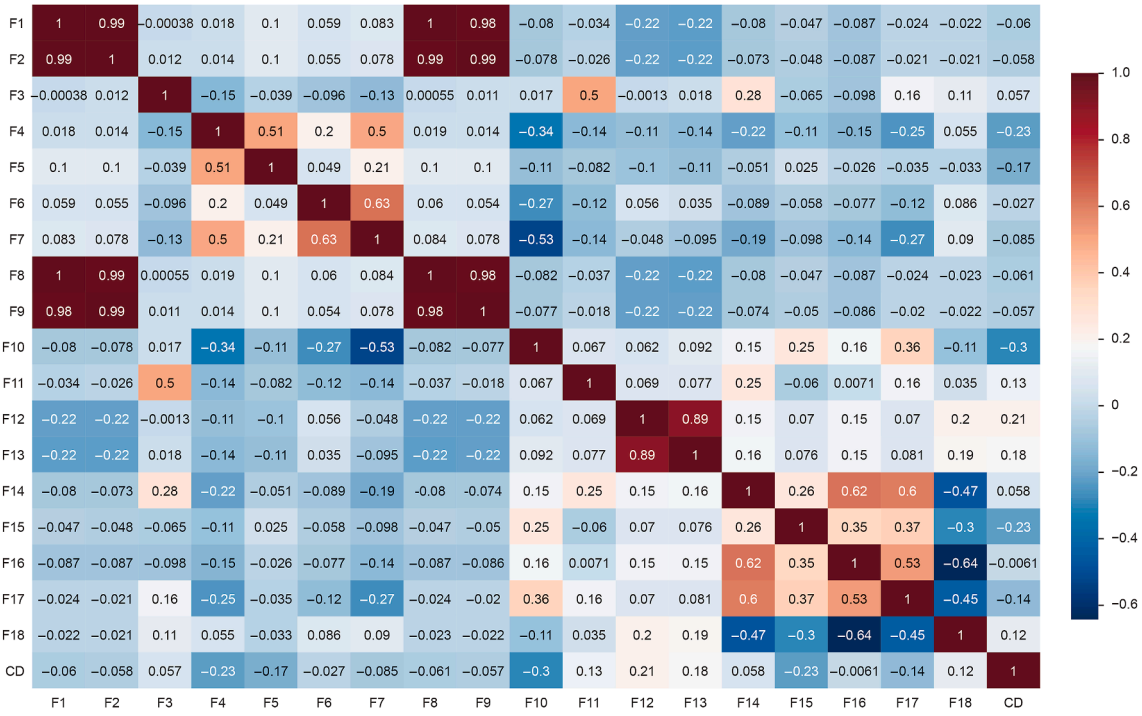

Fig. 8. Kendall correlation coefficient.

Table 4
Ensemble FS scoring scheme results.

Real features	mRMR	RF	F-test	Vote	Mean importance	Final rank
F1	0	0	0	–	–	–
F2	0	0.0001	0	1	0.00003	–
F3	0.020	0.048	0.143	3	0.070	8
F4	0.075	0.049	0.387	3	0.170	5
F5	0	0	0	0	–	–
F6	0.024	0.237	1	3	0.420	1
F7	0.037	0.020	0.621	3	0.226	3
F8	0	0	0.0002	1	0.00006	–
F9	0.0191	0.001	0.022	3	0.0140	12
F10	0.281	0.281	0.436	3	0.332	2
F11	0.020	0.011	0.031	1	0.021	11
F12	0.091	0.078	0.367	3	0.178	4
F13	0.020	0.072	0.363	3	0.152	7
F14	0.023	0.055	0.123	3	0.067	9
F15	0.107	0.101	0.287	3	0.165	6
F16	0.073	0.052	0.034	3	0.053	10
F17	0.001	0	0	1	–	–
F18	0	0	0	0	–	–

algorithms identified casing service duration, casing steel grade, casing wall thickness, and maximum water cut as crucial and pertinent factors in this study, with at least four algorithms selecting them.

The hybrid TriScore-wrapper method (G_1G_2) substantially reduces the number of selected features, typically limiting the final set to an average of 6 features. The results presented in Table 5 illustrate a clear trend, demonstrating that integrating the TriScore approach with the wrapper method (G_1G_2) consistently results in a more concise feature selection. Features are retained based on a voting threshold, requiring each to rank within at least four selection methods. This criterion highlights a strong consensus among the ML algorithms regarding feature importance.

4.4. Model testing final results

Table 6 presents a comparative evaluation of three FS methods, TriScore (G_1), wrapper (G_2), and a Hybrid TriScore-wrapper (G_1G_2), using four boosting-based ML models: AdaBoost, XGBoost, LightGBM, and CatBoost. Performance is assessed based on accuracy, recall, and the number of selected features. The results indicate that the Hybrid TriScore-wrapper (G_1G_2) consistently yields the highest predictive performance with fewer features selected. Specifically, the CatBoost algorithm under this hybrid method achieves the best overall accuracy (0.955) and recall (0.897), using only six features. Notably, across all models, the hybrid method significantly reduces feature dimensionality (from 8 to 13 features in individual methods to just 6 features) while maintaining or enhancing performance metrics. In comparison, both individual methods (TriScore G_1 and wrapper G_2) show relatively similar accuracy and recall values. However, the wrapper method (G_2) exhibits variability in the number of features selected (ranging from 8 to 13 features). The hybrid approach, however, demonstrates superior stability and consensus among models regarding feature selection. These findings highlight the effectiveness of integrating TriScore with the wrapper method, suggesting that the hybrid G_1G_2 method optimizes model performance and interpretability by minimizing feature redundancy and enhancing prediction accuracy.

Fig. 9(a–c) present a comparative evaluation of the predictive performance (accuracy and recall) for three feature selection methods, wrapper (G_2), TriScore (G_1), and the hybrid TriScore-wrapper (G_1G_2), applied across four boosting-based ML models: AdaBoost, XGBoost, LightGBM, and CatBoost. In Fig. 9(a) (wrapper G_2), accuracy values are consistently high (~0.9), while recall is slightly lower across all four models, with CatBoost and XGBoost performing marginally better than AdaBoost and LightGBM. In Fig. 9(b) (TriScore G_1), the accuracy and recall are again consistently strong, though slightly lower than the wrapper method. The gap between accuracy and recall is relatively stable across all

Table 5
The importance of features from three different FS methods.

Feature method	ML models	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	Number of features
TriSFS	All ML models	×	×	✓	✓	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	×	×	12
	Vote	0	0	4	4	0	4	4	0	4	4	4	4	4	4	4	4	0	0	12
Wrapper	AdaBoost	×	×	×	✓	×	✓	✓	×	✓	✓	✓	✓	×	×	✓	✓	×	×	8
	XGBoost	×	×	✓	✓	×	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	13
	LightGBM	×	×	✓	✓	×	✓	✓	×	✓	✓	✓	×	✓	✓	✓	✓	×	×	11
	CatBoost	×	×	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	✓	✓	×	×	×	9
	Vote	0	0	3	4	–	4	4	0	4	4	3	3	–	3	4	3	–	–	10
Hybrid TriSFS-wrapper	AdaBoost	N/A	N/A	N/A	✓	N/A	✓	✓	N/A	N/A	✓	N/A	N/A	N/A	✓	✓	N/A	N/A	N/A	6
	XGBoost				✓		✓	✓			✓				✓	✓				6
	LightGBM				✓		✓	✓			✓				✓	✓				6
	CatBoost				✓		✓	✓			✓				✓	✓				6
	Vote				4		4	4			4				4	4				6

Table 6
Performance metrics of different FS methods along with the four Boosting ML methods.

FS method	Model	Accuracy	Recall	Number of features
TriScore G_1	AdaBoost	0.92	0.823	12
	XGBoost	0.94	0.85	12
	LightGBM	0.935	0.837	12
	CatBoost	**0.944**	**0.852**	**12**
Wrapper G_2	AdaBoost	0.922	0.825	8
	XGBoost	0.943	0.854	13
	LightGBM	0.939	0.838	11
	CatBoost	**0.947**	**0.855**	**9**
Hybrid TriScore-wrapper G_1G_2	AdaBoost	0.943	0.867	6
	XGBoost	0.947	0.88	6
	LightGBM	0.943	0.871	6
	CatBoost	**0.955**	**0.897**	**6**

algorithms, with CatBoost and XGBoost showing marginally superior recall values. Fig. 9(c) (hybrid TriScore-wrapper G_1G_2) exhibits the highest overall performance, demonstrating superior accuracy and recall across all four models. The CatBoost model particularly excels, achieving near-optimal accuracy and the highest recall among the models shown. Importantly, the hybrid method's gap between accuracy and recall is smaller, highlighting balanced predictive effectiveness.

The analysis demonstrates that the six features identified by the hybrid TriScore-wrapper (G_1G_2) method constitute the most informative predictors, as evidenced by increased predictive accuracy and recall after removing redundant or less significant features (Table 6). This indicates that excluded features likely introduced noise rather than meaningful predictive value. Consequently, employing a carefully refined subset of features enhances both the predictive performance and interpretability of the model, underscoring the critical role of rigorous feature selection in developing practical and generalizable ML models.

4.5. SHAP analysis

Fig. 10(a) and (b) provide detailed SHAP values for feature selection results across three distinct methodologies: Wrapper (G_2), TriScore (G_1), and the hybrid TriScore-wrapper (G_1G_2) using CatBoost. Each subplot ranks features according to relative scores, reflecting their importance and selection frequency. Fig. 10(a), representing the wrapper method (G_2), shows a broader distribution of scores, highlighting feature F10 as highly significant, followed by F6 and F4, while multiple features such as F16, F13, and F14 exhibit minimal importance. Fig. 10(b), illustrating the TriScore (G_1) method, similarly identifies feature F10 as the most influential, with F6 and F4 also demonstrating strong relevance.

However, this approach presents additional lower-ranked features (F9, F11, F3) compared to the wrapper method. Fig. 10 (c), the hybrid TriScore-wrapper (G_1G_2), consolidates the outcomes of both approaches, displaying a more refined and concise feature selection set. It reinforces the dominance of feature F10 while retaining only the most consistently high-ranked predictors (F6, F4, F15, F7, F12). Notably, fewer features are maintained in this hybrid approach, indicating increased consensus between the methodologies. The figure generally emphasizes that the hybrid method effectively identifies a smaller, more robust set of critical predictors by integrating the strengths of both the wrapper and TriScore methods. This focused selection is likely to contribute to the improved predictive performance and model interpretability previously documented.

4.6. Discussion

4.6.1. Policy implications

The research findings offer a foundation for constructing models to estimate casing damage demand. They also serve as a reference point for the Gangxi Oilfield, policymakers, and other oilfield companies in their planning. Based on the research findings above, we suggest the following policy recommendations. First, this paper's forecast results show the most influential factors on casing damage. Therefore, this cutting-edge predictive model, which utilizes a novel feature selection process and SHAP analysis, needs to be adopted. This model not only elucidates the most critical factors contributing to casing damage but also enhances decision-making processes by pinpointing specific operational and material characteristics that warrant immediate attention. Policymakers should, therefore, advocate for and facilitate the integration of this predictive tool across the industry.

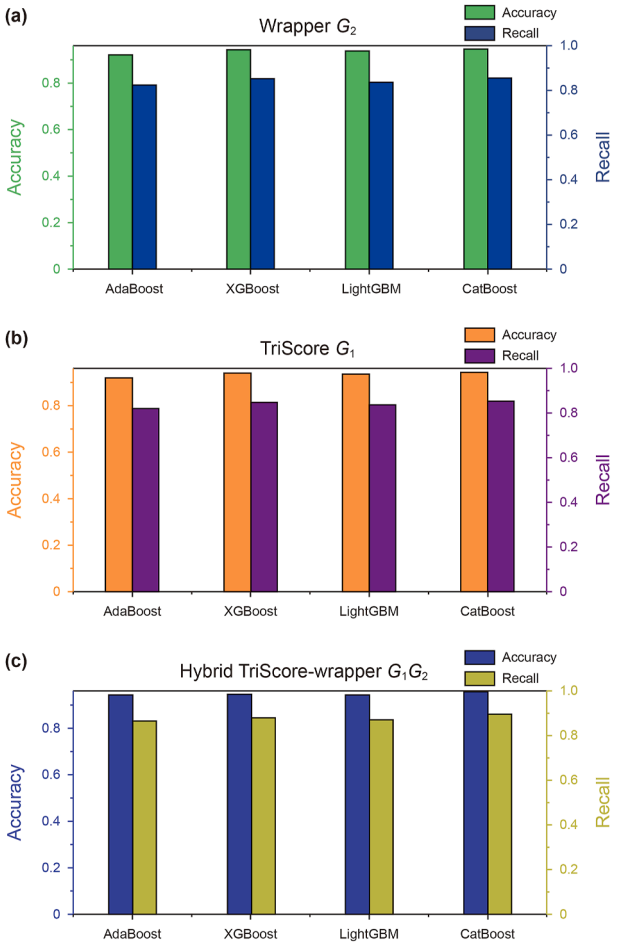

Fig. 9. Evaluation metrics (accuracy and recall) for: (**a**) Wrapper G_2; (**b**) TriScore G_1; (**c**) Hybrid TriScore-wrapper G_1G_2.

Fig. 11(a) and (b)present the temporal evolution of casing damage probability as a function of service time for two representative wells. In Fig. 11(a) (Well 2), the probability of casing damage remains relatively stable (~0.65) during the early service period but rises sharply after approximately 15 years, reaching over 0.85 after 20 years. This suggests a strong age-related degradation pattern, where prolonged exposure to operational stresses likely accelerates the risk of casing failure. In contrast, Fig. 11(b) (Well 1) displays a lower overall probability curve, beginning around 0.15 and gradually increasing. A noticeable inflection point occurs after 16 years, after which the damage probability stabilizes near 0.4. These trends emphasize that casing damage is a cumulative process, with service time as a significant predictor. However, the risk escalation rate and severity may vary depending on construction quality, formation pressure, or operational conditions.

Fig. 12 explores the influence of casing wall thickness on the predicted probability of casing damage across two well categories: damaged and non-damaged. For damaged wells (blue lines), the probability of failure consistently declines as wall thickness increases, with probabilities falling from above 0.70 at 7.0 mm to around 0.50 at 9.5 mm. This inverse relationship confirms the protective role of a thicker casing in resisting mechanical failure.

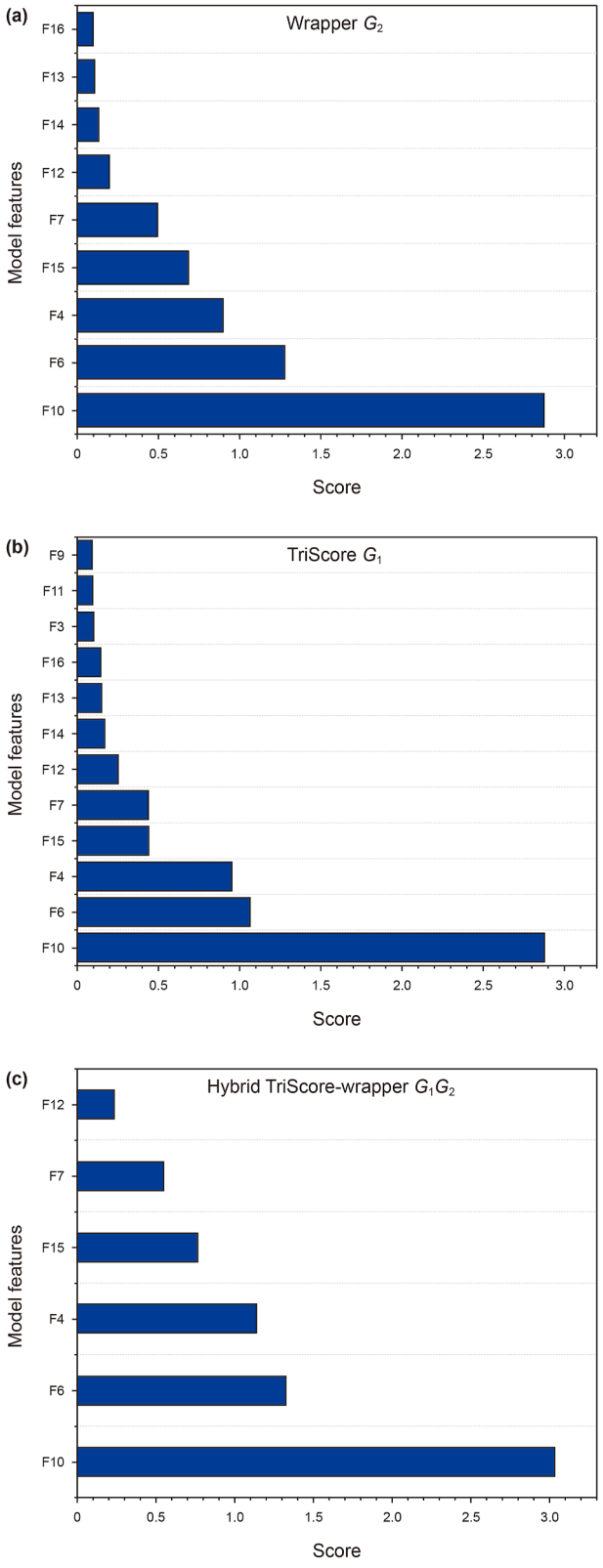

Fig. 10. Shapley analysis for different FS methods using CatBoost.

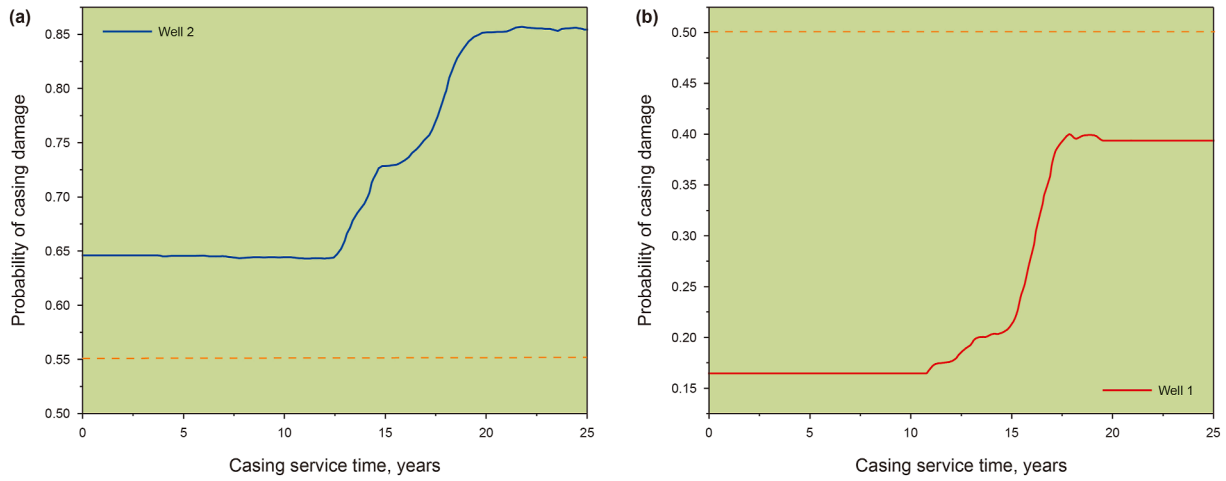

Fig. 11. Sensitivity analysis of casing service time. (**a**) Damaged well, (**b**) non-damaged well.

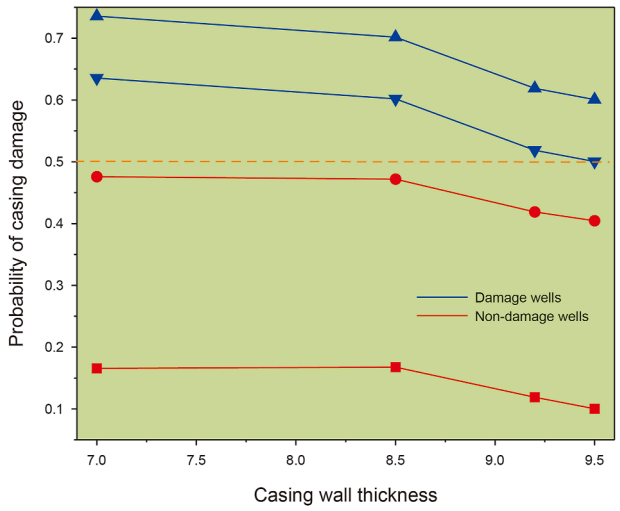

Fig. 12. Sensitivity analysis of casing wall thickness.

For non-damaged wells (red lines), the probability remains significantly lower throughout, further decreasing as wall thickness increases, suggesting that wall thickness contributes to long-term integrity even in initially healthy wells. The clear separation between damaged and non-damaged well profiles across all wall thicknesses supports the conclusion that wall thickness is a critical structural parameter influencing casing longevity, with design optimization offering a viable strategy for mitigating future failure risks.

4.6.2. Limitations of the framework

Our model has shown good generalizability; however, the dataset utilized for training and testing the predictive model comprised 244 oil and water wells, which presents certain constraints regarding data scale. Expanding the dataset to include more wells could enhance the model's accuracy and generalizability by providing a more comprehensive representation of varying operational conditions. Additionally, the quality of the dataset poses notable challenges. Specifically, some fields exhibited missing or aberrant data, necessitating the application of expert judgment and advanced ML techniques for imputation and correction. While these methods help maintain the continuity and usability of the data, they may introduce slight deviations from actual values, potentially affecting the model's predictions. Future data collection and preprocessing improvements may mitigate these issues, leading to more robust and reliable outcomes.

5. Conclusion

This paper presents a novel framework that successfully combines innovative feature selection (FS) approaches with boosting techniques, significantly enhancing the prediction of casing damage. The framework incorporates three FS approaches, TriScore, wrapper, and hybrid TriScore-wrapper, with four interpretable machine learning (ML) models that enhance performance: AdaBoost, XGBoost, LightGBM, and CatBoost. This combination determines the models with the maximum prediction accuracy and assurance of transparency in the prediction process. The approach demonstrates exceptional proficiency in quantifying the comparative significance of features and their influence on casing damage. It achieves a harmonious blend of model reliability and interpretability, augmenting the forecasts' dependability and practicality. This study sets itself apart by considering various engineering, geological, and production aspects and introducing a generalizable TriScore FS approach. It also includes a detailed comparative FS analysis and a SHAP analysis using data from the Gangxi Oilfield.

The study's main findings are as follows.

(1) The CatBoost model, using the hybrid TriScore-wrapper G_1G_2 method, showed excellent predictive performance. Compared to other ML models assessed in the study, it achieved the greatest accuracy (95.5%) and recall rate (89.7%) for casing-damaged wells. The suggested hybrid TriScore-wrapper G_1G_2 FS technique effectively reduces dimensionality without compromising important forecasting accuracy.

(2) The analysis of casing damage causes in the Gangxi Oilfield revealed 12 main elements that influence casing damage. This led to development of a comprehensive set of metrics for evaluating casing damage.
(3) The Bayesian optimization algorithm was creatively utilized to assess the influence of hyperparameters on model performance and to identify the most optimal combination of these parameters.
(4) The assessment of casing damage identified significant factors like the casing service time, casing wall thickness, and perforation density.

While the present model attains high predictive accuracy using routinely logged parameters, its performance could be further strengthened by incorporating direct measurements of buckling and collapse stress, sand-transport velocity, Reynolds number, high-resolution down-hole velocity, and stress-profile logs. Targeted data-collection campaigns aimed at acquiring these variables, particularly during work-over or logging-while-drilling operations, would enable more explicit physics-based descriptors, reduce reliance on proxy features, and refine mechanistic understanding of casing integrity. Integrating such enriched datasets into the proposed workflow constitutes a valuable next step for field diagnostics and model generalizability.

CRediT authorship contribution statement

Juan Li: Resources, Project administration, Funding acquisition, Formal analysis, Data curation. **Mandella Ali M. Fargalla:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Wei Yan:** Supervision, Methodology, Conceptualization. **Zi-Xu Zhang:** Writing – review & editing, Visualization, Validation, Software. **Wei Zhang:** Writing – review & editing. **Zi-Chen Zou:** Writing – review & editing, Visualization. **Tang Qing:** Validation, Resources, Project administration, Investigation, Formal analysis, Data curation. **Tao Yang:** Validation, Resources, Project administration, Funding acquisition, Formal analysis, Data curation. **Chao-Dong Tan:** Validation, Supervision. **Guang-Cong Li:** Writing – review & editing.

Data availability

The data that has been used is confidential.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was funded by the National Natural Science Foundation Project (Grant No. 52274015) and the National Science and Technology Major Project (Grant No. 2025ZD1402205).

Appendix A. Multicollinearity assessment

During our exploratory data analysis, we identified several pairs of highly correlated (multicollinear) features, such as (F1, F2) and (F8, F9). To gauge their effect on model performance and interpretability, we conducted an ablation study where we systematically removed these correlated features and compared results against the complete feature set. Interestingly, rather than improving or leaving accuracy unchanged, removing the correlated features produced a noticeable drop in performance across all four ensemble methods (AdaBoost, XGBoost, LightGBM, CatBoost), as illustrated in Fig. A1(a) and (b).

This outcome supports the position in the literature that tree-based boosting algorithms are robust to multicollinearity and may even derive subtle benefits from partially redundant signals. Consequently, while correlated features might affect interpretability in linear models (e.g., unstable coefficients), they did not impair performance in our experiments, and indeed, retaining them often improved predictive power (Fig. A2(a–d)). We therefore elected to keep these features, guided by both their domain relevance and the evidence that removing them could diminish model accuracy, recall, and F1.

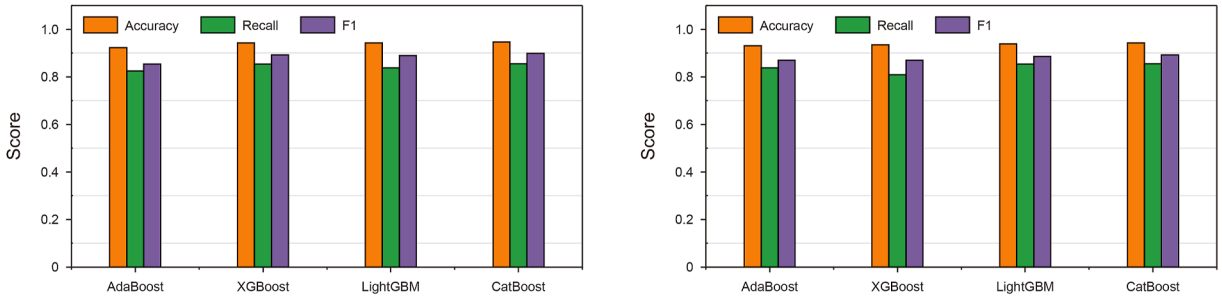

Fig. A1. Multicollinearity assessment.

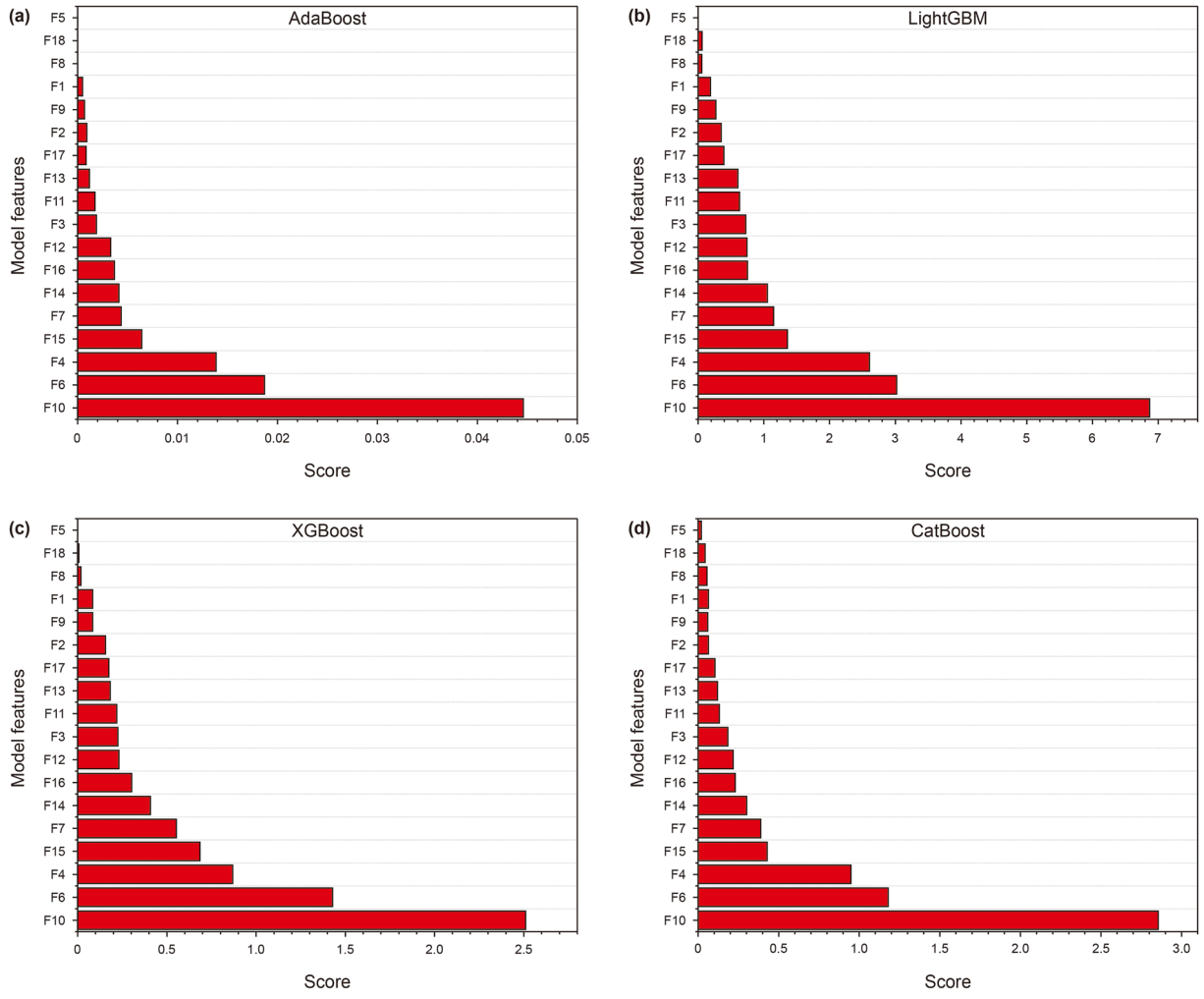

Fig. A2. Feature importance.

References

Abu-Doush, I., Ahmed, B., Awadallah, M.A., et al., 2023. Enhancing multilayer perceptron neural network using archive-based harris hawks optimizer to predict gold prices. J. King Saud Univ. 35 (5), 101557. https://doi.org/10.1016/J.JKSUCI.2023.101557.

Awal, M.A., Masud, M., Hossain, M.S., et al., 2021. A novel bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data. IEEE Access 9, 10263–10281. https://doi.org/10.1109/ACCESS.2021.3050852.

Braik, M., Sheta, A., Kovač-Andrić, E., Al-Hiary, H., et al., 2024. Predicting surface ozone levels in eastern Croatia: leveraging recurrent fuzzy neural networks with grasshopper optimization algorithm. Water, Air, Soil Pollut. 235, 1–40. https://doi.org/10.1007/S11270-024-07378-W/TABLES/17.

Cai, J., Luo, J., Wang, S., et al., 2018. Feature selection in machine learning: a new perspective. Neurocomputing 300, 70–79. https://doi.org/10.1016/J.NEUCOM.2017.11.077.

Chen, Z., Xiao, F., Guo, F., et al., 2023. Interpretable machine learning for building energy management: a state-of-the-art review. Adv. Appl. Energy 9, 100123. https://doi.org/10.1016/J.ADAPEN.2023.100123.

Dai, Y., Yu, W., Leng, M., 2024. A hybrid ensemble optimized BiGRU method for short-term photovoltaic generation forecasting. Energy 299, 131458. https://doi.org/10.1016/J.ENERGY.2024.131458.

Deng, K.H., Zhou, N.T., Lin, Y.H., et al., 2023. Failure mechanism and influencing factors of cement sheath integrity under alternating pressure. Pet. Sci. 20 (4), 2413–2427. https://doi.org/10.1016/J.PETSCI.2023.03.004.

Doush, I.A., Ahmed, B., Awadallah, M.A., et al., 2024. Improving multilayer perceptron neural network using two enhanced moth-flame optimizers to forecast iron ore prices. J. Intell. Syst. 1 (33), 20230068. https://doi.org/10.1515/JISYS-2023-0068/ASSET/GRAPHIC/J_JISYS-2023-0068_FIG_007.

Fargalla, M.A.M., Yan, W., Deng, J., et al., 2024. TimeNet: Time2Vec attention-based CNN-BiGRU neural network for predicting production in shale and sandstone gas reservoirs. Energy 290, 130184. https://doi.org/10.1016/J.ENERGY.2023.130184.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.5555/944919.944968.

Karasu, S., Altan, A., Bekiros, S., et al., 2020. A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. Energy 212, 118750. https://doi.org/10.1016/J.ENERGY.2020.118750.

Li, F., Yan, W., Kong, X., et al., 2024. Study on multi-factor casing damage prediction method based on machine learning. Energy 296, 131044. https://doi.org/10.1016/j.energy.2024.131044.

Li, X., Wang, Z., Yang, C., et al., 2024. An advanced framework for net electricity consumption prediction: incorporating novel machine learning models and optimization algorithms. Energy 296, 131259. https://doi.org/10.1016/J.ENERGY.2024.131259.

Lian, Z., Yu, H., Lin, T., et al., 2015. A study on casing deformation failure during multi-stage hydraulic fracturing for the stimulated reservoir volume of horizontal shale wells. J. Nat. Gas Sci. Eng. 23, 538–546. https://doi.org/10.1016/J.JNGSE.2015.02.028.

Lin, T., Zhang, Q., Lian, Z., et al., 2016. Evaluation of casing integrity defects considering wear and corrosion—Application to casing design. J. Nat. Gas Sci. Eng. 29, 440–452. https://doi.org/10.1016/J.JNGSE.2016.01.029.

Liu, Q., Li, Y., Jiang, H., et al., 2024. Short-term photovoltaic power forecasting based on multiple mode decomposition and parallel bidirectional long short term combined with convolutional neural networks. Energy 286, 129580. https://doi.org/10.1016/J.ENERGY.2023.129580.

Mohamadian, N., Ghorbani, H., Wood, D.A., et al., 2021. A geomechanical approach to casing collapse prediction in oil and gas wells aided by machine learning. J. Pet. Sci. Eng. 196, 107811. https://doi.org/10.1016/J.PETROL.2020.107811.

Noshi, C., Noynaert, S., Schubert, J., 2019. Data Mining Approaches for Casing Failure Prediction and Prevention. International Petroleum Technology Conference (IPTC). https://doi.org/10.2523/IPTC-19311-MS.

Noshi, C.I., Noynaert, S.F., Schubert, J., 2018. Failure predictive analytics using data mining: how to predict unforeseen casing failures. Abu Dhabi International Petroleum Exhibition and Conference. https://doi.org/10.2118/193194-MS.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., et al., 2022. A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics. 2, 927312. https://doi.org/10.3389/FBINF.2022.927312/XML/NLM.

Ren, Y., Lv, Z., Xu, Z., et al., 2023. Slurry-ability mathematical modeling of microwave-modified lignite: a comparative analysis of multivariate non-linear regression model and XGBoost algorithm model. Energy 281, 128143. https://doi.org/10.1016/J.ENERGY.2023.128143.

Shi, X.L., Huang, W.J., Gao, D.L., 2021. Mechanical behavior of drillstring with drag reduction oscillators and its effects on sliding drilling limits. Pet. Sci. 18, 1689–1697. https://doi.org/10.1016/J.PETSCI.2021.09.007.

Song, M., Zhou, X., 2019. A casing damage prediction method based on principal component analysis and gradient boosting decision tree algorithm. SPE Middle East Oil and Gas Show and Conference. https://doi.org/10.2118/194956-MS.

Tang, Q., Deng, J., Wu, H., et al., 2019. Prediction of casing damage in unconsolidated sandstone reservoirs using machine learning algorithms. IEEE International Conference on Computation, Communication and Engineering 185–188. https://doi.org/10.1109/ICCCE48422.2019.9010785.

Theng, D., Bhoyar, K.K., 2024. Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl. Inf. Syst. 66, 1575–1637. https://doi.org/10.1007/S10115-023-02010-5/TABLES/6.

Wang, C., Lin, H., Hu, H., et al., 2024. A hybrid model with combined feature selection based on optimized VMD and improved multi-objective coati optimization algorithm for short-term wind power prediction. Energy 293, 130684. https://doi.org/10.1016/J.ENERGY.2024.130684.

Wang, H.Y., Samuel, R., 2016. 3D Geomechanical modeling of salt-creep behavior on wellbore casing for presalt reservoirs. SPE Drill. Complet. 31, 261–272. https://doi.org/10.2118/166144-PA.

Wang, L., Lv, S.X., Zeng, Y.R., 2018. Effective sparse adaboost method with ESN and FOA for industrial electricity consumption forecasting in China. Energy 155, 1013–1031. https://doi.org/10.1016/J.ENERGY.2018.04.175.

Wang, P., Zhong, C., Fan, S., et al., 2023. Prediction of collapsing strength of high-strength collapse-resistant casing based on machine learning. Processes 11, 3007. https://doi.org/10.3390/PR11103007.

Willson, S.M., Fossum, A.F., Fredrich, J.T., 2003. Assessment of salt loading on well casings. SPE Drill. Complet. 18, 13–21. https://doi.org/10.2118/81820-PA.

Xue, J., 2020. Casing damage classification method using Random Forest algorithms. J Phys Conf Ser 1437, 012131. https://doi.org/10.1088/1742-6596/1437/1/012131.

Yang, S., Han, L., Wang, J., et al., 2021. Laboratory study on casing deformation during multistage horizontal well fracturing in shale gas development and strain based casing design. J. Nat. Gas Sci. Eng. 89, 103893. https://doi.org/10.1016/J.JNGSE.2021.103893.

Yin, F., Shi, B., Huang, G., et al., 2023. Integrity assessment methodology of casing ovality deformation in shale gas wells. Geoenergy Sci. Eng. 224, 211643. https://doi.org/10.1016/J.GEOEN.2023.211643.

Zhang, J., Wu, L., Jia, D., et al., 2022. A machine learning method for the risk prediction of casing damage and its application in waterflooding. Sustainability 14 (22), 14733. https://doi.org/10.3390/su142214733.

Zhang, X., Fan, J., Zou, Y., Sun, W., 2023. Realizing accurate battery capacity estimation using 4min 1C discharging data. Energy 282, 128744. https://doi.org/10.1016/J.ENERGY.2023.128744.

Zhao, Y., Jiang, H., Li, H., 2020. Prediction of casing damage: a data-driven, machine learning approach. International Journal of Circuits, Systems and Signal Processing 14, 1047–1053. https://doi.org/10.46300/9106.2020.14.133.

Zhou, Y., Wang, S., Xie, Y., Zeng, J., et al., 2024. Remaining useful life prediction and state of health diagnosis of lithium-ion batteries with multiscale health features based on optimized CatBoost algorithm. Energy 300, 131575. https://doi.org/10.1016/J.ENERGY.2024.131575.