IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 November 2024, accepted 21 November 2024, date of publication 27 November 2024, date of current version 9 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3507569

RESEARCH ARTICLE

A Novel Multiagent Collaborative Learning Architecture for Automatic Recognition of Mudstone Rock Facies

SAURABH TEWARI[1], **ARVIND PRASAD**[1], **HARSH PATEL**[2], **MUEEN UDDIN**[3], **(Senior Member, IEEE)**, **TAHER AL-SHEHARI**[4], **AND NASSER A. ALSADHAN**[5]

[1]Department of Computer Engineering and Applications, GLA University, Mathura 281402, India
[2]Department of Data Science, Gyan Ganga Institute of Technology and Sciences, Jabalpur 482003, India
[3]College of Computing and Information Technology, University of Doha for Science and Technology, Doha, Qatar
[4]Department of Self-Development Skill, Common First Year Deanship, King Saud University, Riyadh 11451, Saudi Arabia
[5]Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Mueen Uddin (mueen.uddin@udst.edu.qa)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSPD2024R846. The Open Access research is supported by the Qatar National Library QNL.

ABSTRACT Recognizing mud rock lithofacies is essential for mapping the subsurface depositional environments and identifying oil and gas-bearing rock formations. Conventional well logs interpretation techniques are slow, costly and require high domain expertise. Machine learning (ML) techniques have been implemented to automate the recognition of lithofacies from the bulk of well logs generated. However, the reservoir heterogeneity and uneven thickness of rock layers result in imbalanced data conditions that make the ML models biased. This study proposes a novel multiagent collaborative learning architecture (MCLA) to handle the imbalanced data problem during the identification of lithofacies. This research investigates four popular data resampling techniques, i.e. oversampling, SMOTE and ADASYN. Also, resampling techniques are combined with nine different ML classifiers, including Decision tree, ExtraTree, Random Forest, Logistic regression, Support vector machine, K-nearest Neighbour, Naïve Bayes and Ensemble methods. Stacking and voting ensembles combine the outcomes of diverse classifiers working as team members in MCLA. ADASYN, in combination with Stacking, has produced impressive results in terms of accuracy (99.41%) along with MCC (0.98) and G-mean (0.98). The proposed MCLA shows an enhancement of 2% in lithofacies accuracy and an approximately 4% increment in reliability compared with the top-performing Extra Tree classifier considered in this study.

INDEX TERMS Data imbalance, resampling techniques, machine learning, stacked generalization, multiagent collaborative learning.

I. INTRODUCTION

Mudstones are a type of sedimentary rock that contain largely silt and clay-sized particles and play three roles as a hydrocarbon-producing source, cap, and reservoir rock to provide generation, migration, and storage of oil and gas [1]. These rock formations hold sweet spots and shale gas reservoirs that are helpful for oil and gas production. Mudstone formations are complex geological structures that provide a greater challenge to conventional well-log interpretation techniques for identifying subsurface lithofacies layers. The petrophysical properties (such as porosity, permeability, lithofacies, wettability, etc.) are required to be determined accurately for modelling mudstone reservoirs. Recognition of underlying lithofacies is important for the exploration and development phase of oil and gas wells.

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

Reservoir characterization involves the determination of reservoir properties through the data acquired from different geophysical techniques such as core analysis, seismic, and well logs. These geostatic properties are required to develop a trustworthy hydrocarbon reservoir model. However, reservoir properties have an anisotropy property which results in heterogeneous reservoir behaviour. The heterogeneous nature of the reservoir promotes uncertainty in all the sensory data captured during its characterization. The reservoir data are found to be complex with nonlinear, noise, and imbalance problems associated with it. The manual interpretation of recorded well-log data is costly, time-consuming, and tedious work even for expert geologists. Thus, intelligent methods are needed to process the reservoir data for extracting useful information. Several researchers have suggested multiple machine learning (ML) models for extracting facies information from well logs data in conventional reservoirs. However, there is limited research existing for applying these models to unconventional mudstone reservoirs. The modeling of mudstone lithology using ML paradigms is confined to two approaches - unsupervised and supervised classifiers, as reported by several researchers [2], [3], [4], [5]. Mudstone lithology has been comprehensively reviewed by Aplin and Macquaker who also investigated its multiple functions as a source rock, cap rock, and reservoir rock for hydrocarbon production [1]. In 2017, another in-depth examination of mudstone facies in the Henry Mountain Region of Utah was conducted by Li and Schieber [6]. In 2006, Qi and Carr utilized an ANN model to identify carbonate lithofacies in Southwest Kansas using well-log data [2]. Gifford and Agah investigated a multiagent collaborative learning approach to classify the well logs data obtained for the Kansas field in the USA [7]. However, they didn't address the effect of class imbalance occurring in well-log data during the classification task. researcher created a 3-D model of shale facies for the Appalachian basin at the regional scale, using discriminant analysis, ANN, SVM, and fuzzy logic techniques, along with core, seismic, and well-logs data [3]. In 2016, unsupervised cluster analysis was implemented to classify lithofacies and identify productive sweet spots in the Barnett Shale formation [8]. Another study was conducted in which researchers evaluated the effectiveness of both supervised and unsupervised ML models for identifying mudstone facies in the Mahantango-Marcellus and Bakken Shale formations in the United States [5]. Similarly, they also utilized SVM to categorize shale in the Bakken Formation in North Dakota [9]. Homogenous ensemble methods were investigated to identify mud rock layers existing in the Kansas oil and gas field [10]. Further, heterogeneous ensemble methods were also proposed for the recognition of underlying mudstone facies [11]. Gu et al. applied generalized and robust ensemble modelling to classify carbonate diagenetic rock facies [12]. Song et al. [13] applied ensemble based deep learning approach for the estimation of rock porosity using seismic attributes. Liang et al. [14] applied hybrid machine learning model for the identification of lithofacies along with the data mining approach. Laplacian SVM model was utilized to test the semi-supervised approach for identifying underground subsurface facies [15]. Sixteen well logs were taken as input to model the lithofacies and thirteen evolutionary optimizers were used for optimum feature selection [16]. Convolution neural networks (CNN) model was trained on five wireline well logs to identify the facies [17]. Further, a hybrid model was proposed by combining CNN and LSTM models that was trained on image logs data to recognise rock facies [18]. Deep learning model was trained on logging data for 3D modelling of tight sandstone reservoirs [19]. More machine learning models were applied to recognise carbonate sedimentary facies from well logs data [20]. Boosting ensemble methods were implemented to recognise carbonate reservoir facies [21]. Extreme gradient boosting and resampling algorithms were employed to identify lithofacies using well logs data [22]. Further, DNN was also utilized to handle the well logs data related issues and compared with LSTM for identification of lithofacies [23]. The performance of semi-supervised generative adversarial network (SSGAN) for lithofacies identification [24]. Multivariate analysis was also performed on the well logs for change detection to automate the lithofacies identification [25]. Various researcher have tried to rectify the data related issues and tested ML models to classify the well logs data but imbalance issue was rarely reported. More research is required to handle the imbalanced well logs data for the recognition of lithofacies. The performance of ML classifiers is reduced in imbalance well logs data conditions. The classifiers become biased for the majority classes (Classes having large training data samples) while ignoring minority classes (classes having fewer data samples) when these models are trained on imbalanced data. However, only a few have considered the problem of imbalanced data in the petroleum domain associated with lithofacies prediction during the training phase due to the uneven thickness of subsurface rock layers. Initially, imbalance in well logs data was reported in 2018 for the identification of lithofacies [26], [27], [28], [29]. Logging while drilling data was used to identify coal layers using ML models along with two imbalanced data handling techniques via Naïve random oversampling (NRUS) and Synthetic minority oversampling techniques (SMOTE) to improve the classification accuracy [30]. Further, a TwinSVM model was implemented to handle the imbalanced drilling data for the detection of stuck-up pipe conditions [29]. A multi-expert learning system was applied for the recognition of underlying lithofacies and implemented the Reweighting method for compensating the impact of the data imbalance issue [31]. A new a multistage SVM classifier architecture was proposed by modifying the regularization parameter which penalized wrongly classified minority class data points to handle the imbalance well logs issue [32]. ML models was tested to classify different reservoir formation data into seven rock permeability scores [33]. They also reported the imbalance issue in reservoir data and implemented NRUS and SMOTE techniques for handling it [33]. Gaussian mixture model and back-propagation neural network combination was applied to recognize the facies of a tight sandstone reservoir existing in the Sulige gas field [34]. Various hybrid architectures have been proposed to handle the complex data and systems [35], [36]. Therefore, hybrid approach must be tested for handling the imbalance well logs data issue. This paper proposes a novel multiagent collaborative learning architecture (MCLA) for handling data imbalance issues existing in well logs data utilized for lithofacies prediction. In this hybrid architecture, data and algorithm levels are combined using ensemble to achieve higher classification accuracy. Initially, resampling techniques were also implemented to make the data balance, then this modified balance dataset will be given to the Stacking ensemble architecture for classifying the data samples into their respective lithofacies. Also, a comparative study of data imbalance has been performed on well-log data with popular data resampling algorithms in combination with nine ML classifiers for the recognition of lithofacies. A proper preprocessing step has been designed to eliminate the issues related to the data modeling such as noise, high dimensionality, overfitting, and underfitting, model parameter tuning, etc. The impact of data imbalance on classifiers' performance has also been studied by training them on balanced datasets, and imbalanced datasets generated by considering lithofacies of uneven thickness. The primary objectives of this research work are mentioned as given below.

- To design a preprocessing step for eliminating the issues related to the data modeling.
- To study the impact of imbalance in well logs data on supervised classifiers' performance during the recognition of lithofacies,
- To assess the performance of popular resampling techniques for imbalanced well logs data.
- To propose a novel hybrid data-driven architecture for classifying imbalanced well logs data.
- To study the reliability and stability of ML classifiers in the imbalanced logging data condition.
- To study the effects of different well logs acting as controlling variables for lithofacies prediction.
- To review and compare the existing imbalanced data handling approaches.

The remaining paper is organized as follows: Section II provides a brief introduction to data-driven intelligent modeling techniques. Section III describes the experimental evaluation and methodology Section IV discusses the outcomes of this research work. Section V provides the conclusion and future implications.

II. A BRIEF DESCRIPTION OF THE DATA IMBALANCE ISSUE RELATED TO LITHOFACIES

The thickness of subsurface lithofacies layers varies naturally unevenly in their well-log-based measurements. However, when ML-based classifiers have been utilized to identify different rock facies, the imbalanced data condition arises. This imbalanced data refers to a situation in which the distribution of data samples in any class inside the training data is significantly unequal. This means that one class is underrepresented as compared to another, leading to potential bias in the result analysis. Imbalanced data is a common problem in many fields, including healthcare, finance, and fraud detection, among others. The consequences of imbalanced data will be severe. Traditional classification algorithms, which are often optimized for balanced datasets, will struggle to predict the minority class accurately. As a result, models trained on imbalanced data may have poor performance, with high false negative rates for the minority class and low precision and recall overall. This will lead to missed opportunities to detect important phenomena, such as rare diseases or fraudulent activities. Several factors can contribute to imbalanced data, including data collection processes and natural class distributions. For example, in healthcare, rare diseases may naturally have a lower prevalence in the population, leading to imbalanced datasets. Similarly, in fraud detection, the number of fraudulent transactions may be much lower than the number of legitimate transactions, leading to an imbalanced dataset. There are three possible solution approaches for handling imbalanced data issues as given below.

A. DATA LEVEL APPROACH

This approach modifies the training data of classifiers through resampling techniques such as TOMEK [37], Synthetic Minority Over-sampling Technique (SMOTE) [38], Adaptive Synthetic Sampling (ADASYN) [39], etc. to make the number of data samples equal present in each class. The data level approach can be broadly into two types i.e.

1) Undersampling approach. This type of method balances the number of samples in minority and majority classes by removing the data samples existing in majority classes. This approach is not popular because removing the original data samples from the training set reduces the information content and reduces the overall prediction performance of classifiers. These are mostly applied and suitable for problems having large datasets.
2) Oversampling approach: This method tries to make the dataset balance by adding new data samples in minority classes. ADASYN and SMOTE are two popular techniques for addressing the class imbalance in machine learning. Both techniques are designed to generate synthetic samples for the minority class (the class with fewer examples) by interpolating existing examples or creating new ones. This helps to balance the class distribution and improve the performance of the classifier. The main difference between ADASYN and SMOTE lies in how they generate synthetic samples. While SMOTE creates synthetic examples by interpolating between existing minority class examples, ADASYN uses a density

TABLE 1. A brief details of diverse imbalanced data handling techniques.

Ref.	Manipulation Techniques	Key Features
26	Under sampling	1. This approach is straightforward and fast to balance the available dataset. 2. The data points of majority classes are reduced. 3. This method is useful when minority classes have enough data samples available for training. 4. It judges every class separately in multiclass use cases and is preferable in big training data conditions.
26	Oversampling	1. Random over-samplers enhance the quantity of data points in minority classes by duplicating them. 2. It's variants such as SMOTE generate the synthetic data samples for minority classes.
26	Combination of over and under-sampling	1. This approach contains the benefits of over-sampling and under-sampling techniques. 2. Here, data samples of majority classes will be reduced whereas minority classes will be increased. 3. It has limited advantages at data level when compared with ensemble methods.
22	SMOTE Algorithm	1. This popular technique is a variant of oversampling. 2. It produces artificial data points for minority classes.
23	ADASYN Algorithm	1. It also creates synthetic data points for minority classes and minimizes the data related bias error. 2. It generates varying numbers of artificial data for minority classes based on their local data distribution.
21	Tomek link	1. Tomek link is a variant of under-sampling technique that was developed by modifying the Condensed Nearest Neighbour algorithm. 2. It eliminates the majority data samples having lowest Euclidean distance with minority samples.
27	SMOTE-Tomek link combination	1. It combines the benefits of SMOTE and Tomek link techniques. 2. SMOTE creates the artificial data points for minority classes using the k nearest neighbour and Euclidean distance concepts. 3. Similarly, Tomek link technique reduces data samples of majority classes.

distribution to determine the degree of extrapolation between examples. This means that ADASYN focuses more on generating synthetic examples in regions where the density of minority examples is low, making it more effective in dealing with high-dimensional datasets.

Majority Weighted Minority Oversampling Technique is another technique that involves identifying minority class samples that are difficult to learn and informative, and then generating synthetic samples using a weighted version of these samples [38]. Resampling techniques are commonly used to address the class imbalance in data analysis. However, when dealing with highly imbalanced data, these methods have drawbacks. They often involve discarding a significant number of samples from the majority class or generating numerous synthetic samples for the minority class. Such approaches will result in either a loss of valuable information (in the case of undersampling) or an inappropriate increase in the negative correlation between samples (in the case of oversampling). This, in turn, will negatively impact model performance, as highlighted by Wu [40]. Several other popular techniques are also available for resampling, such as TOMEK and its variants, SMOTE variants, etc. The details about the different data manipulation techniques are given in Table 1.

B. ALGORITHM LEVEL APPROACH

This algorithm-level approach aims to modify existing ML methods to address their tendency to favour the majority class. Among these approaches, cost-sensitive techniques are mostly used. These techniques assign a higher cost to the incorrect classification of minority class instances. For example, the cost-sensitive multilayer perceptron (CSMLP) method distinguishes between the importance of class errors using a single cost parameter [44]. Another approach, CLEMS, incorporates a cost-sensitive label embedding technique that considers the relevant cost function [45]. The CS-DMLP model is a deep multi-layer perceptron that employs cost-sensitive learning to regulate a given sample's predicted posterior probability distribution [46]. These approaches typically require expert knowledge to determine the appropriate cost value, which will be challenging to obtain in real-world scenarios [47]. Twin SVM (a variant of the traditional SVM) model was applied for classifying imbalance well-log data [26]. A brief explanation of machine learning models utilized for facies recognition is presented in Table 2.

C. ENSEMBLE APPROACH

Ensemble methods are a type of machine learning approach that blends data-level and algorithm-level techniques to improve accuracy. This is achieved by combining the solutions generated by multiple classifiers [53], [54]. One popular example of such a method is the WEOB2 model developed by Wang et al. in 2015, which leverages online bagging with adaptive weight adjustment to balance the learning bias between the majority and minority classes [55]. Despite these advances, a major challenge in ensemble methods is how to ensure and make use of the diversity of classifiers, which remains an unresolved issue according to Wu et al. 2017 [41]. The details about ensemble classifiers have been given in the section in Table 3. Figure 1 shows different popular approaches utilized for handling data imbalanced issues.

III. METHODOLOGY

In this work, three primary approaches for handling data imbalance have been investigated for the classification of well logs data to identify nine different rock facies. Ten machine learning models have been tested (LR, DT, RF, SVC, XGB, KNN, NB, and ET) along with four popular data manipulation techniques (under-sampling, oversampling, SMOTE, and ADASYN). All the above-described data-related models

TABLE 2. A brief explanation of machine learning models utilized for facies recognition.

Ref.	Machine Learning Models	Key Features
33	Logistic Regression	1. It classifies data samples into respective classes using a probability-based regression approach along with a normal cumulative distribution curve. 2. This model is more suitable for small datasets and assumes that input variables are independent and linearly correlated with output. 3. This model is easy to train. 4. It is highly accurate if the dataset is linearly separable.
34	Decision Tree	1. This model is based on entropy and splits the input data into a tree structure for understanding and prediction. 2. It can be employed for classification, estimation, and feature selection purposes with a high degree of accuracy. 3. It has lesser training time when compared to SVM, ANNs, etc.
35	Support vector classifier	1. It utilises maximum margin boundaries. 2. It is developed for handling high dimensionality of data with the help of kernel trick. 3. It is complex, high training time, hard to find suitable kernel for specific data etc.
36	K- Nearest Neighbour	1. It performs classification operations based on the K nearest neighbors data points. 2. This nonparametric algorithm is easy to implement but requires large storage space. 3. The computation cost of K-NN is high. It takes large time to produce output.
37	Naïve Bayes	1. Naive Bayes classifier uses Bayes theorem and adopts the training variables independent of each other. 2. KNN is more suitable for smaller datasets with ease in application.

TABLE 3. A summary of ensemble methods utilized in this study.

Ref.	Ensemble Models	Key Features
39	Bagging	1. It creates random bootstrap data datasets with replacement from training data and trains the base classifiers with them. 2. The final testing results are determined by the maximum votes acquired from the decisions of the base classifiers. 3. Develop a model having low variance. 4. Chances of overfitting reduce. 5. Risk of overfitting. 6. Maximum and lowest values are ignored in the results
40	Boosting	1. It sequentially trains the base classifiers and updates the weights of wrongly classified data samples in each iteration. 2. The updated weights increase the occurrence of wrongly classified data samples. 3. Develop a model having low bias conditions. 4. Chances of overfitting prevails
41	Random Forest	1. RF will be developed by utilizing decision trees as base learners in Bagging ensemble architecture. 2. Its training time is more the DTs. 3. It will also handle problems related to classification, regression, and feature selection. 4. The chances of overfitting and underfitting are less. 5. It is susceptible to noise and outliers' presence in the training data.
42	Extra Tree	1. This ensemble technique also combines the output of decision trees. 2. It is very conceptually like the Random Forest algorithm except for the way the tree structure is constructed using the Gini index. 3. Normally, the performance of ExtraTree is nearly the same of RF algorithm.
43	Xgboost	1. It combines gradient descent and boosting ensemble in its architecture. 2. This supports parallel processing and takes less training time. 3. It is suitable for classification as well as estimation tasks. 4. It needs categorical features to label encode manually.
11	Voting	1. The training data is divided into random K-fold subsets for training the base classifiers. The yield of these base classifiers is merged using algebraic rules. 2. It improves the individual performance of base classifiers through its architecture. 3. Different base classifiers will also be utilized in its architecture. 4. It is sensitive to noise and outliers present in the training data.
11	Stacking	1. The training data is divided into random K-fold subsets for training the base classifiers. The results of these base-classifiers is merged using a meta-classifier. 2. The individual performance of base classifiers improves in Stacking architecture. 3. Different base classifiers can also be utilized in Stacking architecture. 4. It is sensitive to noise and outliers present in the training data.

have been implemented through Python programming on the Google Collab platform with Scikit learn package version 1.3.0 for machine learning model development and Scikit imbalanced-learn package for data manipulation techniques. This work is done on 4th generation intel core i7-1065G7 processor model configurations along with 8 MB Cache, GHz, four cores, 16 GB RAM, 8 GB graphic cards, and the Windows 10 operating system.

A. DATA DESCRIPTION

The well-log data utilized in this study has been downloaded from the website of the Kansas Geological Survey. This is one of the largest open-source databases containing Geophysical data provided on the internet for research purposes. In this research work, Paradise A well data have been downloaded in Las file format for testing the performance of the different algorithms. This well is situated in the Kansas field of the

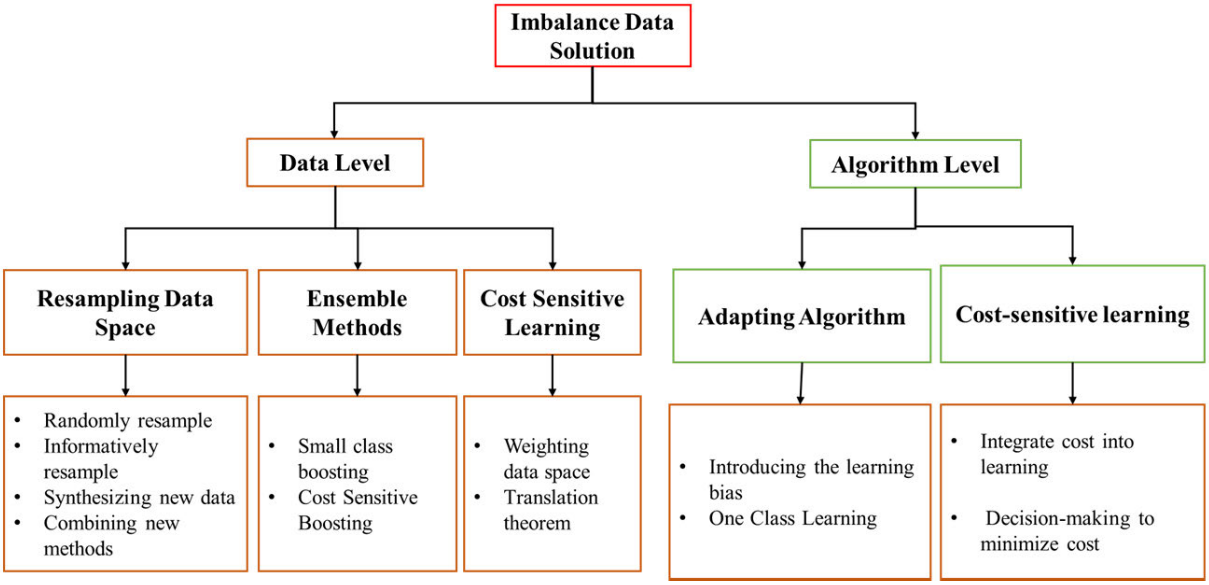

FIGURE 1. Different popular approaches utilized for handling data imbalanced issues.

USA. The data samples having missing values were removed resulting in 2281 data samples made up of 16 different well logs as attributes and nine lithofacies as class labels (Wackstone, Dolomite, Clay, Dolomite Packstone, Dolomite Wackestone, Silt, Packstone, Limestone, Argillaceous clay) as shown in Table 4. The logging data taken from the Kansas Geological Survey website of Deforest well is shown in Figure 2 and the frequency distribution of data samples in each class is shown in Figure 3.

B. DATA PREPROCESSING

Initially, the well-log data was pre-processed to increase the lithofacies identification performance of machine learning models [62]. It cleans the data by removing the samples containing garbage, null, or missing values. The clean data is further normalized to remove the influence of data attributes having large values on smaller ones. The normalization is applied using the formula given below.

$$X_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where xi is the actual data samples, xmin is the minimum value of data samples, and xmax is the maximum value of the data samples. Further, noise filtering was also performed to remove any outliers present in the data that might limit the performance of data-driven machine-learning techniques. Various noise removal filters are available in the literature for well-log data, such as wavelet denoising, moving average, high pass, etc. Savitzky-Golay filter was reported to be suitable for geophysical well logs as it fits a high degree Nth polynomial (where N varies from 5 to 15) trend using least square techniques and convolution technique [63], [64], [65], [66], [67], [68]. This filter has been selected to smoothen the geophysical logging signal without destroying its original characteristics [65]. It is helpful to eliminate the random non-geological noise from the geophysical data [66]. It is designed to match the waveform of a corrupted noisy signal and preserves the height and width of the extrema [68]. The smoothening of geophysical data is standard practice in the petroleum sector. Different commercial software packages for reservoir simulation, well testing, etc., also have similar inbuilt provisions for data smoothening. The Savitzky-Golay filter utilized for denoising the well log data for lithofacies identification is shown in Figure 4.

After the noise removal, the dimensionality of the original data was reduced by eliminating the redundant well logs that were not contributing to the identification of lithofacies. This decreases the size of the original data and the computational cost associated with the machine learning models during facies recognition. Various popular algorithms may be utilized for attribute or feature selection tasks, such as tree-based feature selection, univariate attribute selection, principal component analysis, etc. These paradigms are based on different approaches for selecting attributes or extracting features. Here, the Relief algorithm (RF) was chosen for attribute selection as it focuses on discriminatory information, which is decisive for classification tasks. Most heuristic algorithms assume that attributes are conditionally independent and don't have any interdependencies within the input attributes [69]. However, this assumption does not hold for practical conditions. RF algorithm is aware of the contextual information and correctly estimates the quality of the input attributes, contributing to the prediction and having strong relational dependencies between themselves [69]. RF algorithm family can identify correlation and dependence relationships between the input data attributes. They assign weights and ranks to the logging data attributes based on their contribution to the pattern recognition task for predicting the lithofacies

TABLE 4. The range of variables utilized for lithofacies prediction.

Well logs/Units	Ranges of data	Related Reservoir Property
Depth (DT)/Fts	434-3771	Depth
Sonic logs(SPOR)/pu	2.774-133.212	Formation porosity
Density Log (DEN)	0.65-32.2	Formation density
Gamma-ray log (GR)/API	16.78-414.2	Gamma radiation
Neutron log (NPOR)/pu	0-44.143	Formation Porosity
Density correction log (RHOC0(gm/cc)	0.01-0.296	Formation bulk density
Deep induction log (RILD)ohm m	1.91-65.158	Formation fluid saturation
Medium induction log (RILM)	2.031-140.7	Formation fluid saturation
Deep later log resistivity log (RLL3)	2.866-266.5	Formation fluid saturation
Caliper log 1 (DCAL)/inch	7.468-8.982	Borehole Diameter
Caliper log 2 (MCAL)/inch	6.947-8.9	NA
Microinverse resistivity log (MI)/ohm m	0.609-41.82	Flushed zone resistivity
Micronormal resistivity log (MN)/ohm m	0.138-44.1	Flushed zone resistivity
Acoustic transit time 1 log (TT1)/U s/ft	51.5-235.961	Compaction, stratigraphic, etc.
Acoustic transit time 2 log (TT2)/U s/ft	54.36-662.8	Compaction, stratigraphic, etc.
Spontaneous potential log (SP)/MV	-73.407	Formation thickness and boundaries

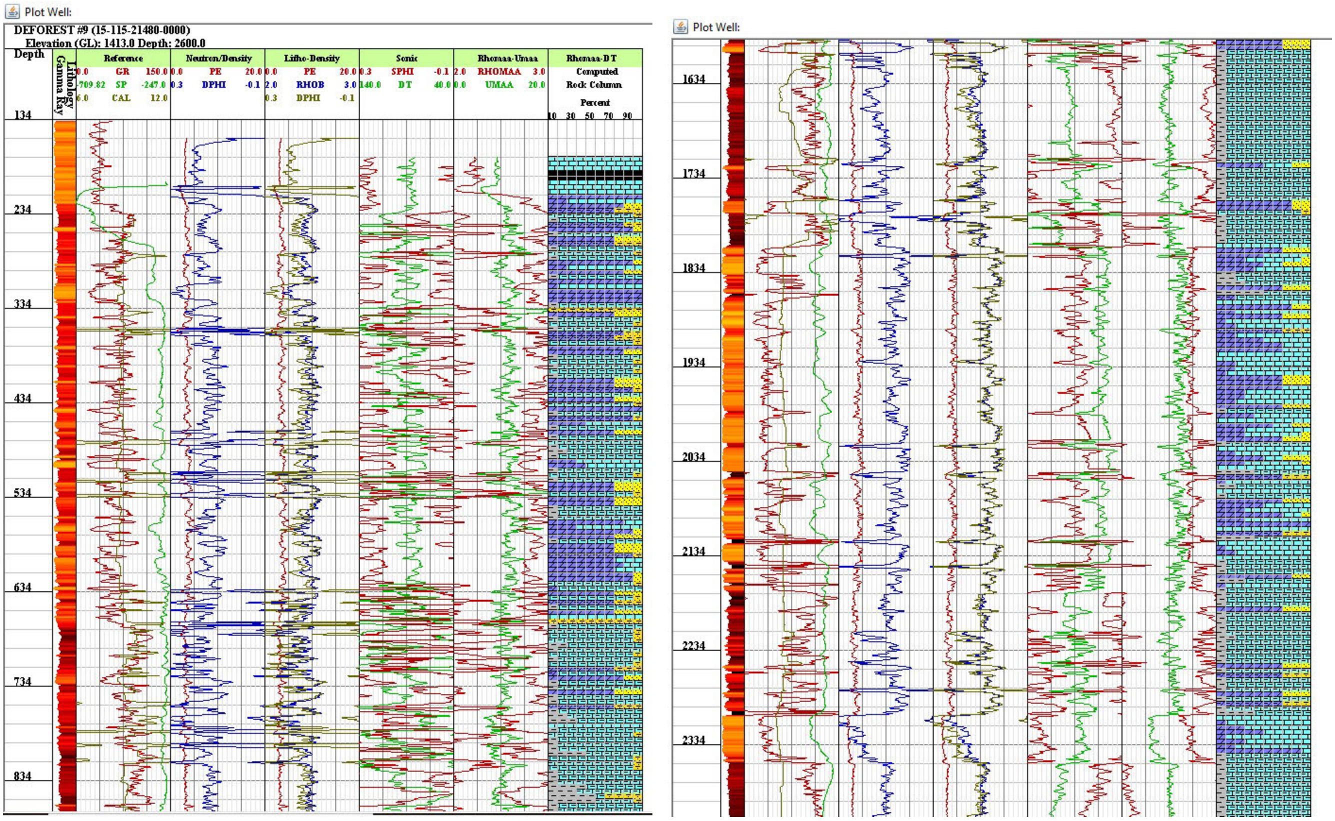

FIGURE 2. The logging data taken from the Kansas Geological Survey website of Deforest well [48].

Figure 5 shows the results of the Relief algorithm on well logs data. The well logs assigned with negative weights were eliminated as they didn't contribute to the identification of lithofacies. SP, GR, DT, RILD, DPOR, GR, TT, and NPOR were given positive weights and recognized as crucial well-logs. After selecting the data attributes, the well-log data were partitioned into training and testing sets using the K-fold cross-validation technique. This technique splits the data randomly into K subsets. Here, (K-1) subsets were used to train the machine learning models, whereas the Kth subset was used to test the trained model. This step was recurring until all the subsets acted as testing subsets. Further, it averages the accuracies of the machine learning model obtained in the iterations. The cross-validation technique saves the model from overfitting and underfitting conditions. Additional test results have been provided in the Appendix I section of the manuscript from Table 16-19 to compare the impact of different cross-validation techniques on lithofacies datasets. The K-fold cross-validation technique is primarily used in this research work to avoid overfitting and underfitting conditions. Since there is a significant difference between the number of samples in majority and minority classes.

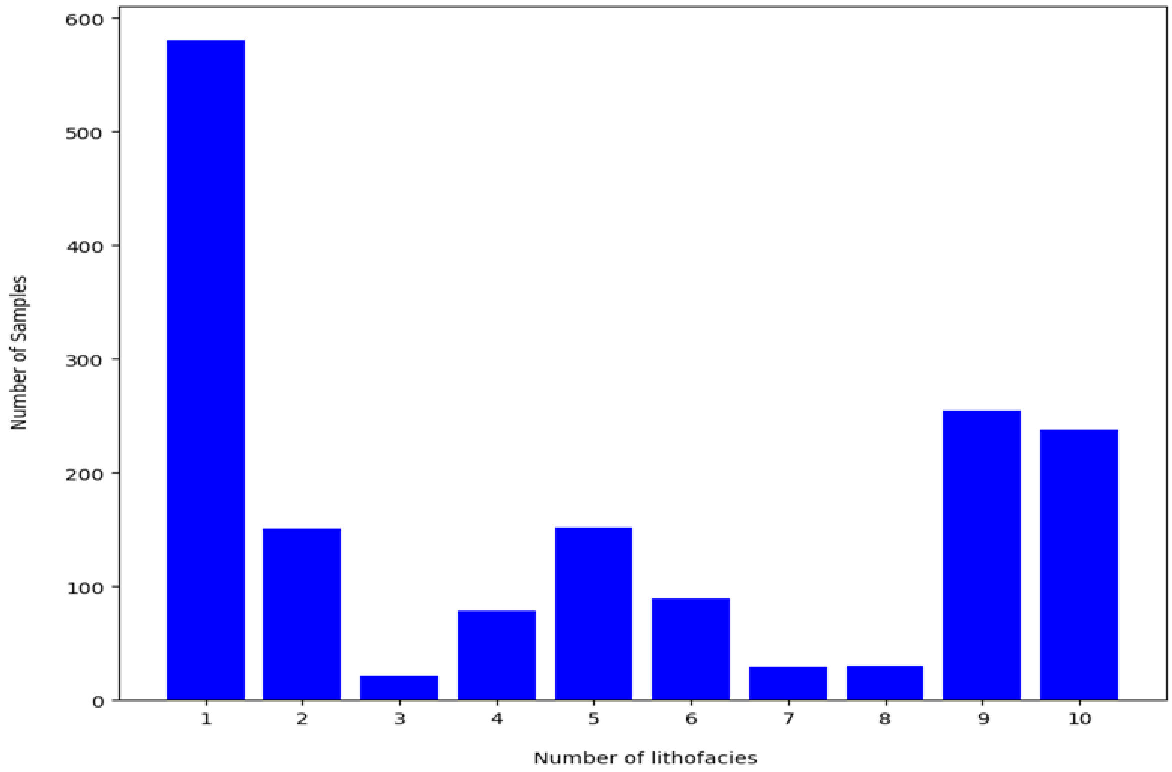

FIGURE 3. The frequency distribution of data samples in each class.

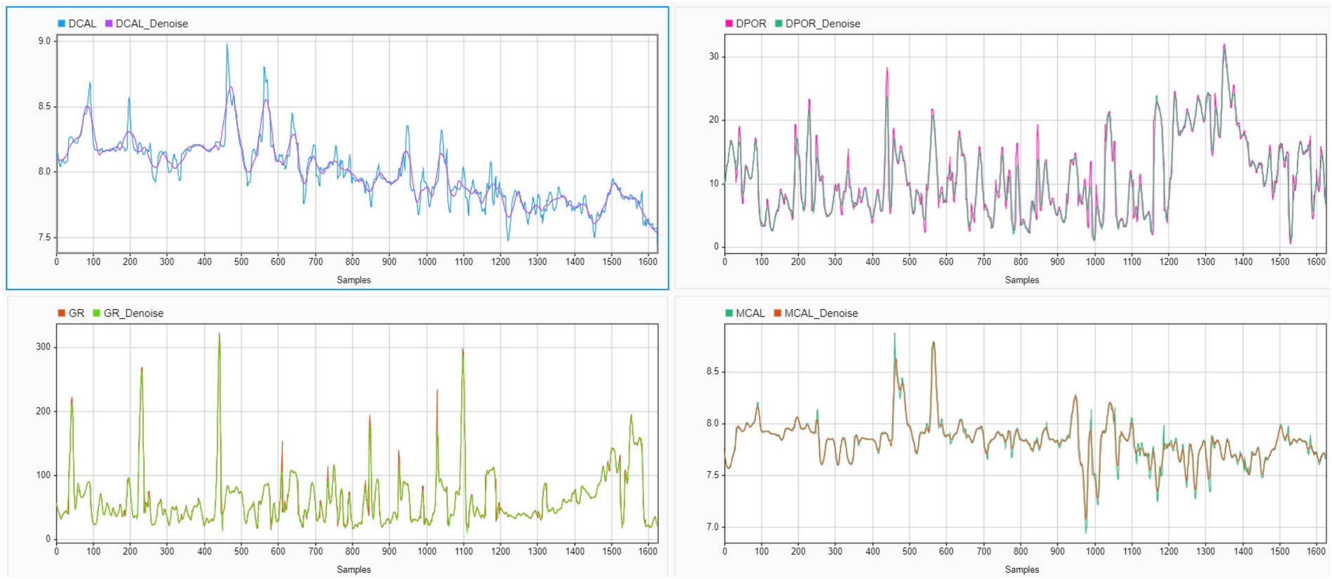

FIGURE 4. The Savitzky-Golay filter utilized for denoising the well log data for lithofacies identification.

It is quite difficult for Stratified k-fold cross-validation or any other cross-validation techniques to divide into k-folds so that each fold has approximately the same proportion of samples from each class as the original data. Therefore, this task is separately handled by data resampling techniques considered in this study. The random_state parameter was set at 42, as per the parameter range recommended by the Scikit learn library. The parameters of machine learning models were also optimized using Bayesian Optimizer to provide the best possible results. This ensures the good performance of the trained model on unseen datasets and is quite helpful for generalization. Machine learning models are prone to overfitting and underfitting conditions during training and testing operations which make the model useless. Learning

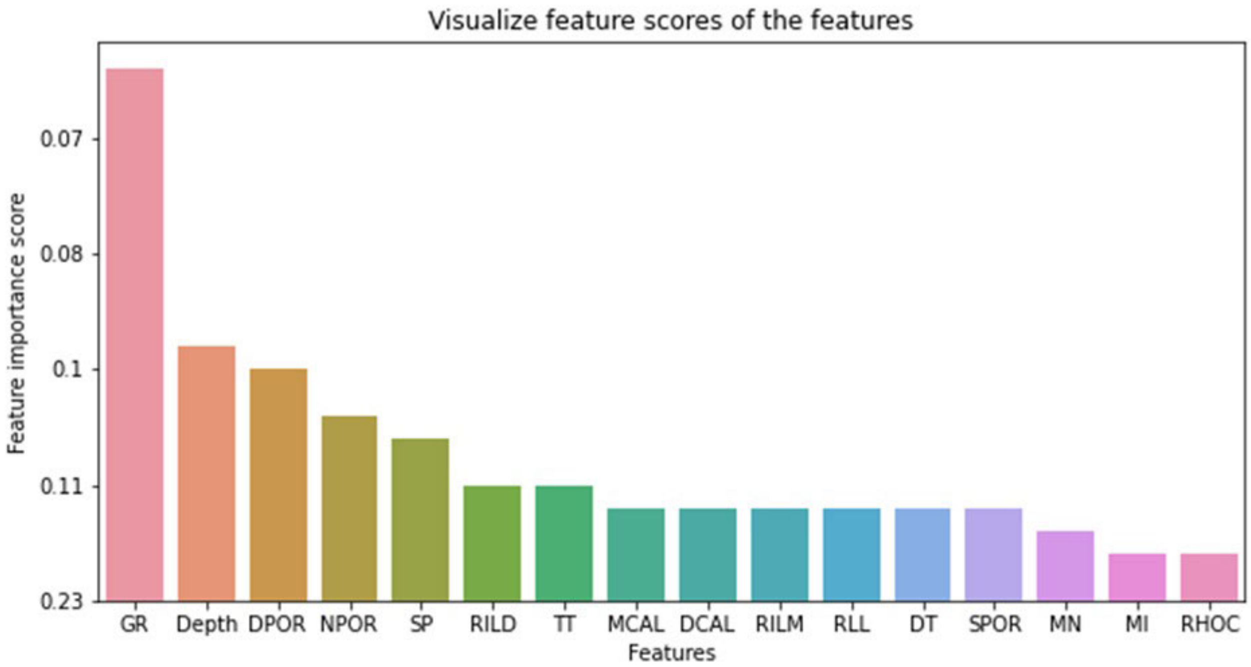

FIGURE 5. The well logs arranged in the descending order of their importance weights.

curves were generated to minimize and identify the feasible search ranges of model parameters to reduce overfitting and underfitting conditions. Figure 7 shows the learning curves generated for four different classifiers. The stable search range does not have any abrupt variations in the training and validation scores. The Bayesian optimizer searches in this stable range to find the optimal values for model parameters. Figure 6 shows the optimization iteration of the DT model for lithofacies prediction. Table 5 shows the search range of each machine-learning model obtained from validation curves and the optimal value of model parameters.

Figure 7 shows the learning curves generated for four different classifiers. The stable search range does not have any abrupt variations in the training and validation scores. The Bayesian optimizer searches in this stable range to find the optimal values for model parameters.

C. PERFORMANCE INDICATOR VARIABLES

The performance of optimally tuned machine learning models was investigated with six performance indicator parameters viz—Matthew correlation coefficient (MCC), Precision, recall, accuracy, f1 scores, and G-mean. MCC and G-mean are performance indicator parameters specially utilized for imbalanced data conditions. In case of data imbalance, classification accuracy, precision, recall, and F1 score are unreliable criteria for performance assessment [70], [71], MCC and G-means are newer and popular parameters to evaluate the classification performance of ML models [70], [71]. MCC considers true and false positives and negatives while calculating the accuracy for classes of different sizes and provides balanced measures [71]. MCC will become zero if any true and false positive and negative values are zero [70]. It is reported that the performance of MCC remains consistent and reliable in case of imbalanced data conditions when compared to accuracy and F-1Score [70]. G-mean is a performance indicator parameter that measures the capability of the classifier model to balance precision and recall [72]. It is maximum when both the values are equal. It's a better performance indicator when compared to accuracy [73]. Therefore, MCC and G-mean are additionally calculated along with other parameters as given below.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + F_P} \tag{2}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{3}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{4}$$

$$F1 - score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \tag{5}$$

where accuracy is a popular performance measurement metric for classifiers, Tp is true positive cases, FP is false positives, TN is true negatives, and FN is false negatives. F1score values have been calculated to maintain the authenticity of precision and recall values. These parameters may be influenced by the data imbalance condition and show good results. To avoid biased result conditions, MCC and G-means are estimated to maintain the reliability of these parameters.

$$MCC = \frac{T_P \cdot T_N - F_P \cdot F_N}{\sqrt{(T_P + F_P) \cdot (T_P + F_N) \cdot (T_N + F_P) \cdot (T_N + F_N)}} \tag{6}$$

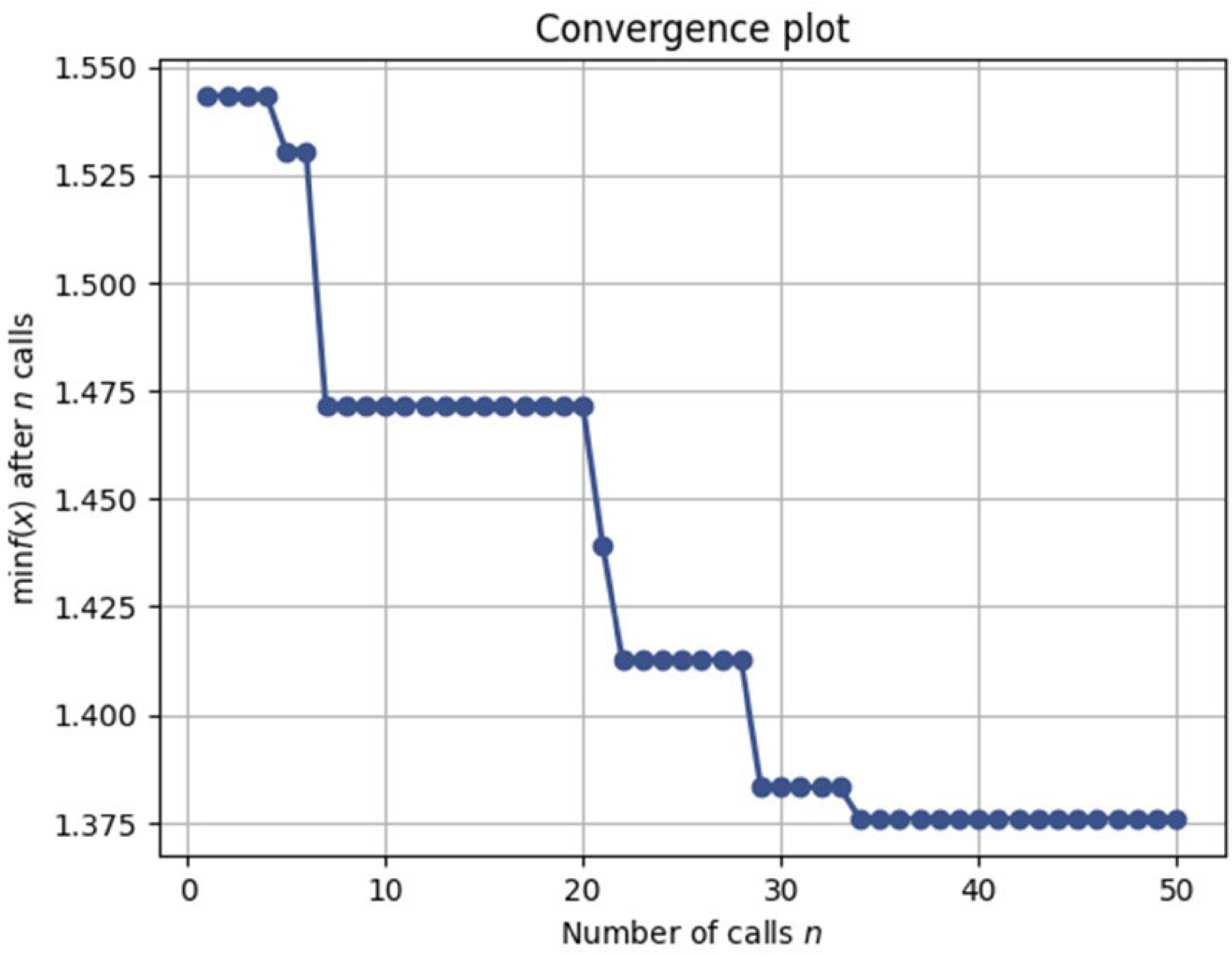

FIGURE 6. Bayesian optimization during training of GB model for lithofacies classification.

TABLE 5. The range of variables utilized for lithofacies prediction.

Machine learning models	Model variables	Search Range	Optimum values settings
Random Forest (RF)	N_estimators Max_depth Criterion	1-500 1-100 gini", entropy", log_loss"	250 32 entropy
Gradient Boosting	N_Estimators Learning_rate Random_state	1-1000 0-1 Jan-50	100 0.1 42
SVM	Class_weight Gamma Kernel Probability	dict or 'balanced' Scale, auto 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' Bool,False	Balanced Scale rbf TRUE
KNNs	N_neighbors Weights	1-100 Uniform, distance	25 uniform
Logistic regression	Multi_class Solver Max_iter Random_state	'auto', 'ovr', 'multinomial' 'lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky', 'sag', 'saga' 1-100000 int, RandomState instance	Multinomial Lbgfs 10000 int
Stacking	Base classifier Meta Classifier	LR, DT, RF,SVC,XGB,KNN,NB,ET LR, DT, RF,SVC,XGB,KNN,NB,ET	DT+RF+ET+XGB+ADA+GRAD ADA+GRAD

where PCK is the iteration for the K class that has been predicted, TCK is the iteration for the K class that has been correctly predicted, SC is the correctly classified data points, and TS is the number of training data samples. MCC considers TP, TN, FP and FN to calculate the performance of the classification models [70], [71], [72]. It can be

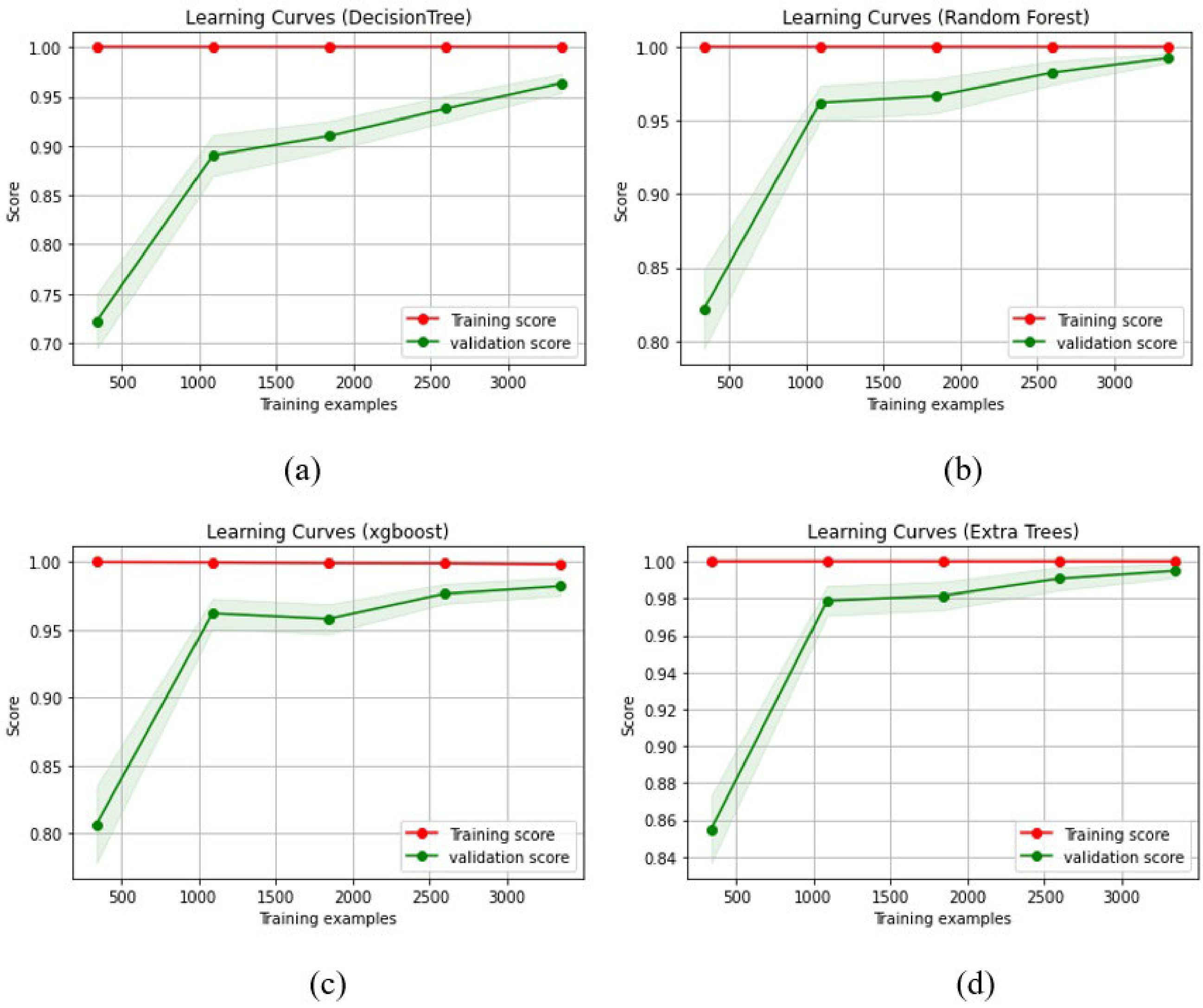

FIGURE 7. Learning curves of tree-based algorithms. (a) Decision Trees, (b) Random forest, (c) Xgboost, and (d) Extra Trees.

directly estimated from the confusion matrix. Its value ranges between the interval of -1 to 1, with 1 showing agreement with classification models and 100% accurate results, -1 showing complete disagreement with the classification model, and 0 meaning predictions are uncorrelated [72]. MCC, along with G-mean, is not affected by the data imbalance condition [72], [73], as given below.

$$G - mean = \sqrt{(T_{\mathrm{P}} * T_{\mathrm{N}})} \tag{7}$$

where TPrate and TNrate are true positive and negative rates. Higher values of the MCC and G-means are essential for ensuring the authenticity and reliability of good classification results [73]. The G-means range from 0 to 1. The classification results with a G-means value nearer to 1 are considered good. G-means and MCC utilise TP, TN, FP, and FN to calculate the model performance, making them more reliable in data imbalance, whereas the F1score includes only three terms: TP, FP, and FN. The conceptual framework of the MCLA architecture is shown in Figure 8.

IV. RESULTS AND DISCUSSION

This section investigates the result generated while testing different machine learning models on the well-log dataset. Eight supervised classifiers were initially trained and tested without any data resampling technique, as shown in Table 6. This will help to understand the impact of adding a resampling technique with the classifier models. The performance of each classifier was tested on well logs data and evaluated using metrics such as accuracy, precision, recall, f1 score, MCC, and G-mean. Table 6 shows that ET has the highest accuracy (100, 98.76) among all other models; however, it has lower MCC and G-mean values (0.97, 0.96) due to imbalance. Similarly, RF and XGB have lower MCC and G-mean values. LR, SVC, and KNN have secured zero G-mean values, indicating that they are not fit for data

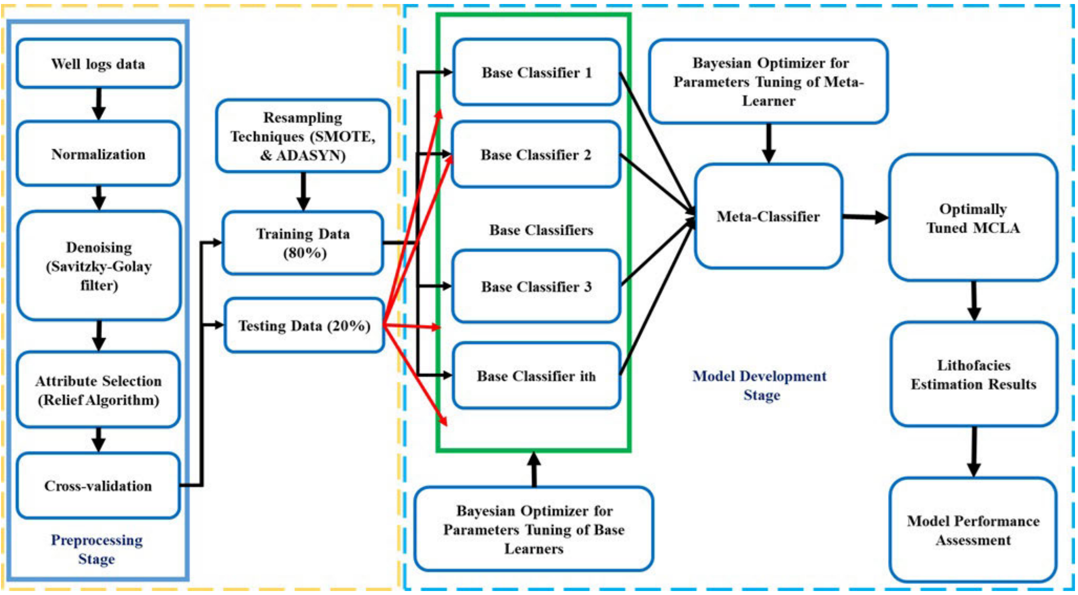
Well logs data
Normalization
Denoising (Savitzky-Golay filter)
Attribute Selection (Relief Algorithm)
Cross-validation
Preprocessing Stage
Resampling Techniques (SMOTE, & ADASYN)
Training Data (80%)
Testing Data (20%)
Base Classifier 1
Base Classifier 2
Base Classifiers
Base Classifier 3
Base Classifier ith
Bayesian Optimizer for Parameters Tuning of Base Learners
Bayesian Optimizer for Parameters Tuning of Meta-Learner
Meta-Classifier
Model Development Stage
Optimally Tuned MCLA
Lithofacies Estimation Results
Model Performance Assessment

FIGURE 8. The conceptual framework of the MCLA architecture.

TABLE 6. Classification results of well logs data with different ML models without any data resampling techniques.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	69.88	64.32	0.36	0.35	0.32	0.55	0
DT	100	90.15	0.87	0.88	0.87	0.87	0.88
RF	100	97.01	0.98	0.92	0.95	0.96	0.91
SVC	79.82	75.04	0.69	0.57	0.6	0.68	0
XGB	100	96.3	0.97	0.93	0.94	0.95	0.91
KNN	73.1	67.83	0.41	0.41	0.38	0.59	0
NB	65.53	60.45	0.53	0.54	0.5	0.52	0.45
ET	100	98.76	0.97	0.96	0.98	0.97	0.96

TABLE 7. Classification results of well-log data with different ML models with random undersampling technique.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G – Mean
LR	58.12	55.08	0.37	0.49	0.38	0.46	0
DT	100	65.02	0.54	0.65	0.56	0.57	0.63
RF	100	75.92	0.67	0.77	0.7	0.7	0.75
SVC	62.5	55.36	0.4	0.52	0.42	0.46	0
XGB	100	72.75	0.64	0.75	0.67	0.66	0.73
KNN	43.12	33.56	0.3	0.33	0.26	0.26	0
NB	61.25	55.18	0.46	0.5	0.47	0.47	0
ET	100	74.69	0.68	0.81	0.73	0.69	0.8

imbalance conditions, but the rest of the performance parameters have higher values. This suggests that in imbalance cases other parameters are not fully trustable. These classifiers were showing slight fluctuations even in the MCC and G-mean values due to the unequal number of data samples in each class or target value.

Firstly, the random undersampling technique was utilized to generate the balanced dataset for training purposes. This technique removes the extra data samples from the majority classes. Table 7 shows the performance of various ML classifiers in combination with the random undersampling technique. When compared with the individual performance of classifiers recorded in Table 6, the classifiers show a decrease in the classification performance due to the lesser number of data samples available in training data when random undersampling was applied. This condition arises due to the random under-sampling technique, which eliminates the extra data points from the majority classes and reduces the data available to learn the hidden pattern.

In the next step, the random oversampling technique was tested to understand its impact on the classification performance of ML classifiers considered in this study. It is observed from Table 8 that the value of performance indicator parameters increases significantly due to the addition of duplicate data samples in the minority classes. The random oversampling technique increases the data samples in the minority classes by adding the duplicate data samples existing within each class. This doesn't add new information to the training data but makes the classifier overfit on the data samples because the same samples may occur in testing data despite cross-validation. The G-mean and MCC values of

TABLE 8. Classification results of well logs data with different ML models with random oversampling technique.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	68.65	65.55	0.51	0.65	0.53	0.58	0.62
DT	100	88.22	0.89	0.85	0.87	0.85	0.84
RF	100	96.13	0.98	0.94	0.95	0.95	0.93
SVC	90.99	83.65	0.77	0.84	0.79	0.79	0.83
XGB	99.89	94.02	0.96	0.93	0.94	0.92	0.92
KNN	91.34	78.2	0.68	0.85	0.73	0.74	0.84
NB	60	56.41	0.49	0.51	0.47	0.47	0.48
ET	100	97.71	0.99	0.96	0.97	0.97	0.95

TABLE 9. Classification results of well logs data with different ML models with SMOTE resampling technique.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	73.64	59.92	0.47	0.59	0.49	0.52	0.55
DT	100	88.4	0.86	0.85	0.84	0.85	0.83
RF	100	98.06	0.98	0.93	0.95	0.97	0.91
SVC	91.23	82.42	0.78	0.83	0.79	0.78	0.82
XGB	99.83	95.43	0.95	0.92	0.93	0.94	0.91
KNN	92.46	77.32	0.72	0.82	0.75	0.72	0.81
NB	68.52	57.46	0.51	0.55	0.5	0.49	0.5
ET	100	98.94	0.99	0.97	0.98	0.98	0.96

LR, SVC, and NB in Table 8 improve when compared with Table 7.

Further, the SMOTE resampling algorithm was implemented to enhance the number of data samples existing in each class. The SMOTE stands for synthetic minority oversampling technique, which introduces synthetic data samples for minority classes by generating new data points in the segment line and joining two original data points existing in the feature space [38]. The number of synthetic data points generated depends upon the requirement of the minority class. Accordingly, the data points of K-nearest neighbours were chosen [38] to balance the training data. This SMOTE oversampling overcomes the disadvantage of the random oversampling technique as it generates newer synthetic data points to avoid overfitting conditions during the testing phase. SMOTE was tested with eight different classifiers to identify the lithofacies using imbalanced well-log data, as shown in Table 9. It is observed that SMOTE slightly enhanced the individual performance of certain classifiers in terms of accuracy, precision, recall, and F1 score but not for MCC and G-mean values if Table 9 and Table 6 are compared. These values don't enhance synthetic data samples, indicating no improvement in strong classifiers such as ET, XGB, and RF performance.

ADASYN is another popular resampling technique that generates synthetic data samples for minority classes to generate balanced training datasets. ADASYN standards for adaptive synthetic sampling approach enhance learning by data distribution among minority classes. It also tries to reduce the bias error introduced due to data imbalance and adaptively manipulate the classification boundary towards the data samples rigid for classification [39]. ADASYN was tested on the well logs data in combination with diverse classifiers, as mentioned in Table 6. The performance of strong classifiers such as RF, XGB, and ET doesn't show any noticeable improvement as no additional information content has been added in the training set. The performance of weak classifiers such as KNN, LR, DT, and SVC improves significantly, as shown in Table 10, compared with Table 6.

Ensemble methods are one of the most popular multi-base classifier algorithms designed to enhance performance. These methods comprise several techniques that are utilized for the classification of lithofacies. Initially, two approaches, bagging and boosting, were investigated to understand the influence of data imbalance. Bagging utilizes two main techniques, i.e. bootstrapping and aggregating, to train the weak base classifiers. It randomly samples the training data with replacement [57]. When the same base classifiers are used in any ensemble architecture, it is known as the homogeneous ensemble method [10], [11]. Table 11 shows the Bagging ensemble architecture in combination with different base classifiers. However, no significant improvement was observed with the Bagging architecture. SVC and KNN combination failed to maintain their G-mean score, as shown in Table 11 and became unreliable. However, their performance became more reliable and stable than earlier performances when boosting ensemble architecture was utilized, as shown in Table 12. This also indicates that boosting algorithms maintain the stability and reliability of classification results more consistently, even with critical imbalance conditions, as observed in Table 12. This may be a key to adding stability to the MCLA. However, a powerful resampling technique must be integrated with an ensemble approach to balance the training data; otherwise, they will not exceed their performance.

Further, Voting and Stacking ensembles were used to enhance the performance of base classifiers in the heterogenous architecture approach. These ensembles were also combined with the resampling techniques to balance the training datasets. It was observed that this MCLA became more stable and reliable, as shown in Tables 13 and 14. However, the performance of the voting classifier is inferior to that of

TABLE 10. Classification results of well logs data with different ML models with ADASYN resampling technique.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	72.3	67.31	0.54	0.63	0.57	0.6	0.6
DT	100	90.15	0.91	0.9	0.9	0.87	0.89
RF	100	96.66	0.98	0.94	0.96	0.95	0.93
SVC	92.65	87.69	0.84	0.87	0.85	0.84	0.86
XGB	99.78	94.55	0.96	0.91	0.93	0.93	0.91
KNN	93.51	82.6	0.78	0.89	0.82	0.79	0.88
NB	56.81	59.05	0.54	0.55	0.52	0.5	0.5
ET	100	98.24	0.99	0.96	0.97	0.97	0.97

TABLE 11. Classification results of well logs data with different ML models with Bagging ensemble technique.

Modified Ensembles	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
BG + DT	100	92.61	0.93	0.86	0.88	0.9	0.84
BG + SVC	79.26	76.44	0.73	0.54	0.57	0.69	0
BG + KNN	74.43	72.75	0.59	0.48	0.49	0.64	0
BG + RF	99.52	94.37	0.96	0.88	0.91	0.92	0.86
BG+ET	100	96.65	0.97	0.9	0.92	0.94	0.89

TABLE 12. Classification results of well logs data with different ML models with Boosting ensemble technique.

Modified Ensembles	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
ADA	100	91.5	91.4	90.7	0.92	0.9	0.9
GB	100	94.37	0.95	0.91	0.92	0.92	0.9
XGB	100	95.95	0.97	0.95	0.95	0.94	0.94
LGB	100	96.3	0.97	0.94	0.95	0.95	0.93

TABLE 13. Classification results of Voting ensemble architecture for lithofacies identification.

Modified Ensembles	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
SMOTE +LR+DT+RF+SVC+ KNN+NB+XGB [HARD VOTING]	92.51	86.29	0.9	0.76	0.86	0.82	0.67
SMOTE+LR+DT+RF+SVC+ KNN+NB+XGB [SOFT VOTING]	98.57	90.68	0.9	0.84	0.86	0.88	0.82
ANASYN+[LR+DT+RF+SVC+ KNN+NB+XGB [HARD VOTING]	99.31	92.97	0.91	0.94	0.93	0.91	0.93
ANASYN+LR+DT+RF+SVC+ KNN+NB+XGB [SOFT VOTING]	99.81	94.02	0.96	0.95	0.94	0.92	0.94

TABLE 14. Classification results of well logs data with different ML models with Stacking ensemble technique.

Modified Ensembles	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
SMOTE+[DT+RF+ET+XGB +XGB(META)]	100.0	97.89	0.98	0.98	0.98	0.97	0.97
ADASYN + [ADA +DT+RF+ET+XGB +XGB(META)]	100.0	99.41	0.98	0.99	0.98	0.98	0.98

stacking due to the additional meta-classifier algorithm layer used to combine outcomes of base classifiers in Stacking. It is also observed that decision tree-based algorithms are better suited for accuracy enhancement in MCLA. However, when combined with boosting algorithms and resampling techniques, this architecture became more stable as the MCC, and G-mean values improved accordingly. LR, KNN, NB, and SVC were also tried as base classifiers, but no significant improvement was observed in the classification results. The highest performance was achieved when ADASYN combined with the Stacking ensemble (ADASYN + [ADA+DT+RF+ET+XGB +XGB(META)]), forming the core of MCLA architecture as shown in Table 14. The performance of ADASYN was observed to be higher than SMOTE because ADASYN focused on generating synthetic data points in minority classes where classification is complex and adapting according to the challenging aspects of imbalance well logs data. The results shown in Tables 14 and 15 clearly show the superiority of multiagent Stacking MCLA for handling imbalanced data. Table 15 also shows the comparative analysis of the proposed approach with the existing research works to establish its supremacy.

The proposed MCLA architecture contains provisions to control data-related issues, imbalanced data, and classification. It was made using a heterogenous ensemble model, representing a committee of experts from diverse

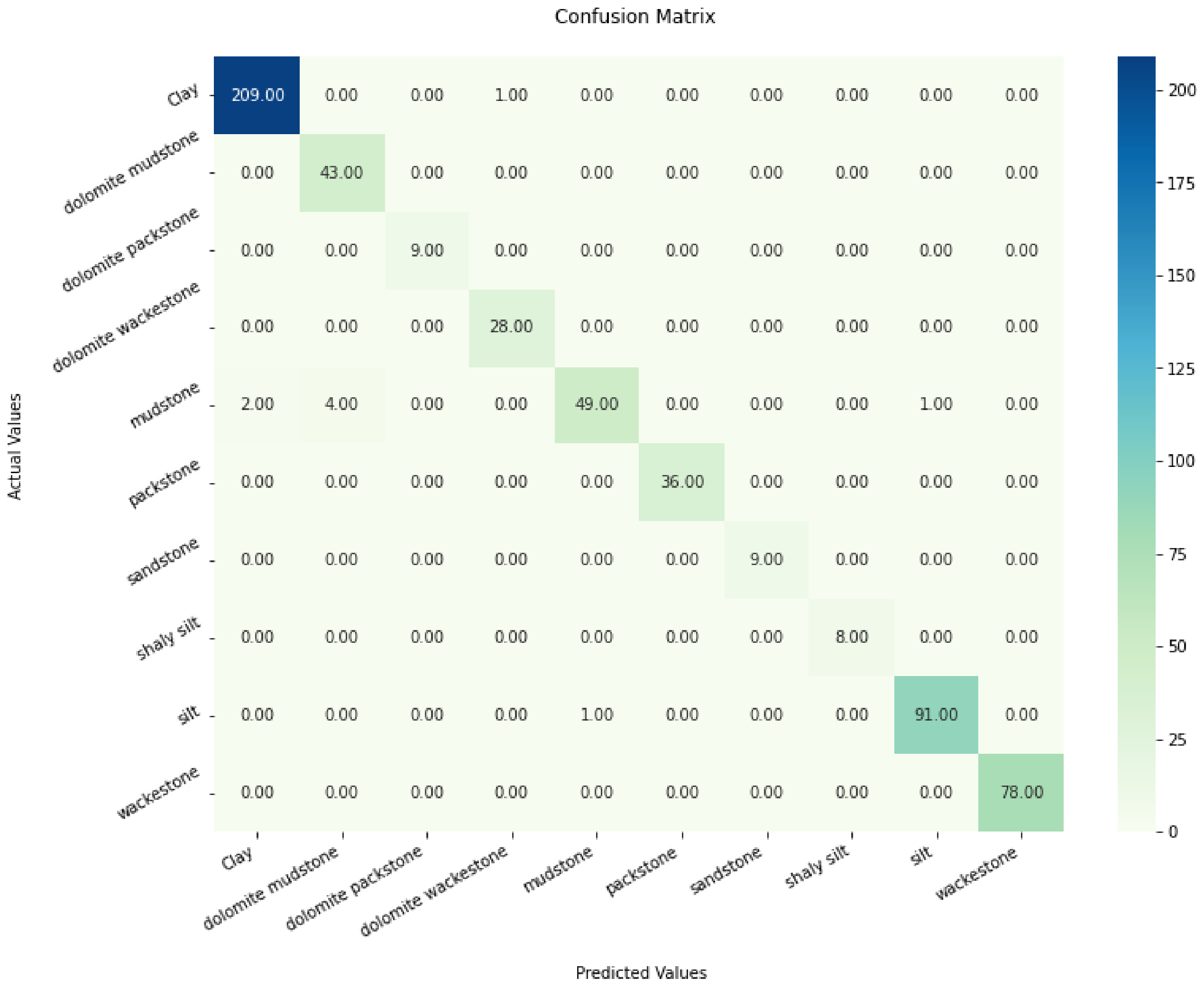

FIGURE 9. The confusion matrix for test data samples of multiagent Stacking architecture during the lithofacies identification.

domains, ensuring better classification accuracy. Due to ensemble architecture at its core, MCLA easily exceeds the individual performance of base classifiers. Different popular algorithms have been tested to classify mudstone lithofacies in the Kansas region. The MCLA architecture optimally combines several techniques (ADASYN + [ADA +DT+RF+ET+XGB+XGB(META)]) at its core, achieving the highest possible classification accuracy and stability even for critical imbalance datasets. Collaboration between different classifiers has been achieved using a heterogeneous ensemble approach, i.e., Stacking and Voting. During the training of base classifiers, the ensemble approach also provided an additional advantage of splitting the imbalance data internally into balanced subsets using the cross-validation technique. This splitting of training well logs data into smaller subsets, which helps to reduce data bias, which in turn reduces prediction error and enhances overall classification accuracy. The presence of ADASYN in the MCLA architecture also lowers bias errors in two ways. Firstly, ADASYN reduces the bias error introduced by imbalance through improvement in learning. The improvement in learning is achieved by generating difficult minority-classes data samples through weighted distributions that are hard to learn [59]. Secondly, the decision classification boundary should be shifted towards the complex generated data samples. The reduction of bias error can be monitored by enhancing values of precision and recall parameters. The bias error can also be minimized by increasing the complexity of modelling through the hybridization with a multi-agent approach in MCLA architecture. The presence of more training features and the size of training samples also minimize the bias error in MCLA. The variance error is reduced by integrating cross-validation, feature selection and ensemble methods in MCLA. Tables 19 and 20 of the appendix section show additional classification results of Voting and Stacking ensembles based on approximately 50,000 training data samples extracted from five different oil and gas wells situated near the Deforest well initially taken into

TABLE 15. Classification results of Voting ensemble architecture for lithofacies identification.

Research work	Data Type	Models	Max Training Accuracy %	Max Testing Accuracy %	Max Precision	Max Recall	Max F1 score	Max MCC	Max G - Mean
Proposed Approach	Well logs data	MCLA	100	99.41	0.98	0.99	0.98	0.98	0.98
Liang et al. [14]	Well logs	GWO-SVM	100	95.2055	NA	NA	NA	NA	NA
Zhong et al. [20]	Logging while drilling data	LR, SVM, ANNs, RF,XGB, SMOTE and NROS	NA	NA	0.72	0.8	0.75	NA	NA
Tsuchihashi et al. [21]	Drilling data	3D-CNN	NA	NA	0.3941	0.6527	NA	NA	NA
Jiang et al. [22]	Well logging data	Adaptive Multiexpert Learning	NA	94	NA	NA	NA	NA	NA
Dutta et al. [23]	Well logs data	Multistage SVM	NA	89.09	NA	NA	NA	NA	NA
Gifford, A. Agad [7]	Well logs data	Multiagent learning	NA	81.5	NA	NA	NA	NA	NA

consideration in this research work. The confusion matrix for test data samples of multiagent Stacking architecture during the lithofacies identification is shown in Figure 9.

The real-field implementation of ML models also comes with various challenges and limitations, such as technical, organisational, and ethical aspects. Gathering data in real-time is often challenging due to the lack of intelligent sensors in the mechanical oil drilling rigs. Data is often scarce, incomplete, noisy, or unstructured, so training the ML models becomes difficult. Oil and gas companies rarely share their data for research purposes. Limited open-source oil and gas databases allow researchers to test their newer technologies. ML technology is resource-intensive, has high computational power requirements, and demands extensive data storage facilities. Deployment and scaling of these ML models in the oil and gas field to compute a bulk amount of data in real-time scenarios is challenging as it has large energy requirements, challenges while integrating with the existing systems, continuous ML model maintenance and monitoring requirements, issues of data privacy, needs protection from adversarial attacks, involvement of high-cost, and ethical considerations. High domain expertise, strategic planning, high monetary funds, and a focus on ethical and responsible AI practices are required to address the abovementioned limitations. Researchers and engineers have already overcome similar challenges and limitations in several megaprojects with team efforts, and ML models will soon be deployed in the form of intelligent drilling rigs, which are under development phase in several research groups.

V. CONCLUSION

This study investigates the multiagent collaborative learning approach for the identification of lithofacies. The number of training data samples belonging to each class, as shown in Figure 3, indicates the uneven distribution of data samples in training data. The proposed architecture is quite helpful in handling imbalanced data conditions. It is clear from the analysis of results reported in Tables 6-14 that the proposed architecture has shown promising improvements in classification performance and maintained its stability with imbalanced well-logs data. Diverse resampling techniques and different classifiers have been tested to investigate the classification results' reliability using MCC and G-mean parameters. Under-sampling and oversampling techniques were found to be poor while handling data imbalances, whereas SMOTE and ADASYN had impressive results. The results of Table 13-14 clearly show that ADASYN outperformed SMOTE because ADASYN reduces the bias error related to data. ADASYSN, in combination with Stacking, has produced impressive results in terms of accuracy (99.41%) along with MCC (0.98) and G-mean (0.98). However, SMOTE has given a slightly lower classification result when compared with ADASYSN in Stacking, such as accuracy (97.89%), MCC (.97), and G-mean (0.97), which are also remarkable. The heterogeneous combination of decision tree-based base classifiers has consistently performed satisfactorily, while boosting algorithms as base classifier members have added reliability to the classification results. The collaboration between decision tree-based algorithms was found to be beneficial in terms of classification accuracy in the Stacking and voting. The high variation in the number of data samples was maintained in the training data utilized for investigating the performance of MCLA. Despite the high degree of data imbalance, the diverse classifiers combined through Stacking in MCLA have produced satisfactory results for recognising lithofacies. The highly diverse classifier members are essential to maintain the robustness of MCLA such that each member complements their performance and ensures the overall performance of the architecture. However, there is a trade-off between the team members section and team size, as well as the operation time and the decision. Computational cost, time,

TABLE 16. Classification results of well logs data with different ML models without any data resampling techniques along with K-fold cross-validation (Same as Table 6).

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	69.88	64.32	0.36	0.35	0.32	0.55	0
DT	100	90.15	0.87	0.88	0.87	0.87	0.88
RF	100	97.01	0.98	0.92	0.95	0.96	0.91
SVC	79.82	75.04	0.69	0.57	0.6	0.68	0
XGB	100	96.3	0.97	0.93	0.94	0.95	0.91
KNN	73.1	67.83	0.41	0.41	0.38	0.59	0
NB	65.53	60.45	0.53	0.54	0.5	0.52	0.45
ET	100	98.76	0.97	0.96	0.98	0.97	0.96

TABLE 17. Classification results of well logs data with different ML models without any data resampling techniques along with stratified cross-validation.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	69.6	68.71	0.59	0.43	0.43	0.6	0
DT	100	88.4	0.86	0.85	0.85	0.85	0.84
RF	100	95.43	0.98	0.93	0.95	0.94	0.92
SVC	77.27	75.92	0.67	0.77	0.68	0.7	0.75
XGB	100	96.3	0.97	0.93	0.95	0.95	0.92
KNN	72.53	70.65	0.5	0.46	0.46	0.62	0
NB	61.64	63.44	0.58	0.57	0.54	0.55	0.52
ET	100	98.24	0.99	0.98	0.99	0.97	0.98

TABLE 18. Classification results of well logs data with different ML models without any data resampling techniques along with leaving one out cross-validation.

Classifiers	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean
LR	69.6	68.71	0.59	0.43	0.43	0.6	0
DT	100	89.63	0.86	0.86	0.86	0.86	0.85
RF	100	95.95	0.98	0.93	0.95	0.94	0.92
SVC	77.27	75.92	0.67	0.77	0.68	0.7	0.75
XGB	100	96.3	0.97	0.93	0.95	0.95	0.92
KNN	72.53	70.65	0.5	0.46	0.46	0.62	0
NB	61.64	63.44	0.58	0.57	0.54	0.55	0.52
ET	100	98.59	0.99	0.99	0.99	0.98	0.98

TABLE 19. Classification results of Voting ensemble architecture for lithofacies identification. (approximate 50,000 data samples extracted from five oil and gas wells situated near Deforest well initially taken).

Modified Ensembles	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean	Execution Time (Sec)
SMOTE + LR + DT + RF + SVC + KNN + NB + XGB [HARD VOTING]	94.3	88.44	0.95	0.0.80	0.89	0.85	0.7	10.38
SMOTE + LR + DT + RF + SVC + KNN + NB + XGB [SOFT VOTING]	98.9	93.8	0.95	0.89	0.88	0.9	0.85	10.92
ANASYN + [LR + DT + RF + SVC + KNN + NB + XGB [HARD VOTING]	99.44	94.97	0.95	0.95	0.94	0.94	0.96	9.99
ANASYN + LR + DT + RF + SVC + KNN + NB + XGB [SOFT VOTING]	99	96.4	0.97	0.96	0.95	0.94	0.96	10.01

TABLE 20. Classification results of Stacking ensemble architecture for lithofacies identification. (approximate 50,000 data samples extracted from five oil and gas wells situated near Deforest well initially taken).

Modified Ensembles	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score	MCC	G - Mean	Execution Time (Sec)
SMOTE + [DT + RF + ET + XGB + XGB (META)]	100.0	98	0.99	0.99	0.98	0.98	0.98	11.78
ADASYN + [ADA + DT + RF + ET + XGB + XGB (META)]	100.0	99.01	0.99	0.99	0.99	0.99	0.99	12.65

team member size, team member type, data complexity, and hardware availability are challenges and constraints hindering MCLA development. Before training these models, a proper preprocessing layer has been utilized to eradicate problematic data samples. This is also challenging as the data from different oil and gas fields may contain different data-related issues. Diverse deep learning models need to be tested as team members in the MCLA to develop a more robust architecture. Such models may reduce the dependency on the preprocessing step. A transfer learning-based model will be proposed and trained for the well-log datasets from different oil and gas fields to replace MCLA. This may overcome various limitations and challenges related to MCLA, such as computational cost, time, dependency on the preprocessing step, etc. The proposed MCLA model may be tested for diverse field datasets having similar severe imbalances and data-related issues.

APPENDIX

See Tables 16–20.

REFERENCES

[1] A. C. Aplin and J. H. S. Macquaker, "Mudstone diversity: Origin and implications for source, seal, and reservoir properties in petroleum systems," *AAPG Bull.*, vol. 95, no. 12, pp. 2031–2059, Dec. 2011.

[2] L. Qi and T. R. Carr, "Neural network prediction of carbonate lithofacies from well logs, big bow and sand Arroyo creek fields, southwest Kansas," *Comput. Geosci.*, vol. 32, no. 7, pp. 947–964, Aug. 2006.

[3] G. Wang and T. R. Carr, "Methodology of organic-rich shale lithofacies identification and prediction: A case study from Marcellus shale in the appalachian basin," *Comput. Geosci.*, vol. 49, pp. 151–163, Dec. 2012.

[4] F. Anifowose, J. Labadin, and A. Abdulraheem, "Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines," *Appl. Soft Comput.*, vol. 26, pp. 483–496, Jan. 2015.

[5] S. Bhattacharya, T. R. Carr, and M. Pal, "Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus shale, USA," *J. Natural Gas Sci. Eng.*, vol. 33, pp. 1119–1133, Jul. 2016.

[6] Z. Li and J. Schieber, "Detailed facies analysis of the upper cretaceous Tununk shale member, Henry mountains region, Utah: Implications for mudstone depositional models in epicontinental seas," *Sedimentary Geol.*, vol. 364, pp. 141–159, Feb. 2018.

[7] C. M. Gifford and A. Agah, "Collaborative multi-agent rock facies classification from wireline well log data," *Eng. Appl. Artif. Intell.*, vol. 23, no. 7, pp. 1158–1172, Oct. 2010.

[8] A. Avanzini, P. Balossino, M. Brignoli, E. Spelta, and C. Tarchiani, "Lithologic and geomechanical facies classification for sweet spot identification in gas shale reservoir," *Interpretation*, vol. 4, no. 3, pp. SL21–SL31, Aug. 2016.

[9] S. Bhattacharya, P. K. Ghahfarokhi, T. R. Carr, and S. Pantaleone, "Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the Marcellus shale, north America," *J. Petroleum Sci. Eng.*, vol. 176, pp. 702–715, May 2019.

[10] S. Tewari and U. D. Dwivedi, "Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs," *Comput. Ind. Eng.*, vol. 128, pp. 937–947, Feb. 2019.

[11] S. Tewari and U. D. Dwivedi, "A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies," *J. Petroleum Explor. Prod. Technol.*, vol. 10, no. 5, pp. 1849–1868, Jun. 2020.

[12] Y. Gu, Y. Li, Y. Yang, B. Xiao, D. Zhang, and Z. Bao, "Classification pattern of lacustrine carbonate diagenetic facies and logging-based data-driven prediction via a generalized and robust ensemble learning: A demonstration of pre-salt profile, Santos basin," *Geoenergy Sci. Eng.*, vol. 223, Apr. 2023, Art. no. 211543.

[13] J. Song, M. Ntibahanana, M. Luemba, K. Tondozi, and G. Imani, "Ensemble deep learning-based porosity inversion from seismic attributes," *IEEE Access*, vol. 11, pp. 8761–8772, 2023.

[14] H. Liang, C. Yun, M. J. Kan, and J. Gao, "Research and application of element logging intelligent identification model based on data mining," *IEEE Access*, vol. 7, pp. 94415–94423, 2019.

[15] Z. Li, Y. Kang, D. Feng, X.-M. Wang, W. Lv, J. Chang, and W. X. Zheng, "Semi-supervised learning for lithology identification using Laplacian support vector machine," *J. Petroleum Sci. Eng.*, vol. 195, Dec. 2020, Art. no. 107510.

[16] D. A. Wood, "Optimized feature selection assists lithofacies machine learning with sparse well log data combined with calculated attributes in a gradational fluvial sequence," *Artif. Intell. Geosci.*, vol. 3, pp. 132–147, Dec. 2022.

[17] J. Kim, "Lithofacies classification integrating conventional approaches and machine learning technique," *J. Natural Gas Sci. Eng.*, vol. 100, Apr. 2022, Art. no. 104500.

[18] J.-J. Liu and J.-C. Liu, "Integrating deep learning and logging data analytics for lithofacies classification and 3D modeling of tight sandstone reservoirs," *Geosci. Frontiers*, vol. 13, no. 1, Jan. 2022, Art. no. 101311.

[19] F. Liu, X. Wang, Z. Liu, F. Tian, Y. Zhao, G. Pan, C. Peng, T. Liu, L. Zhao, K. Zhang, S. Zhang, X. Liu, and R. Zhao, "Identification of tight sandstone reservoir lithofacies based on CNN image recognition technology: A case study of Fuyu reservoir of Sanzhao sag in Songliao basin," *Geoenergy Sci. Eng.*, vol. 222, Mar. 2023, Art. no. 211459.

[20] X. Hou, P. Lian, J. Zhao, Y. Zai, W. Zhu, and F. Wang, "Identification of carbonate sedimentary facies from well logs with machine learning," *Petroleum Res.*, vol. 9, no. 2, pp. 165–175, Jun. 2024.

[21] W. J. Al-Mudhafar, M. A. Abbas, and D. A. Wood, "Performance evaluation of boosting machine learning algorithms for lithofacies classification in heterogeneous carbonate reservoirs," *Mar. Petroleum Geol.*, vol. 145, Nov. 2022, Art. no. 105886.

[22] D. Zheng, M. Hou, A. Chen, H. Zhong, Z. Qi, Q. Ren, J. You, H. Wang, and C. Ma, "Application of machine learning in the identification of fluvial-lacustrine lithofacies from well logs: A case study from Sichuan basin, China," *J. Petroleum Sci. Eng.*, vol. 215, Aug. 2022, Art. no. 110610.

[23] S. Q. Khan and F. U. D. Kirmani, "Applicability of deep neural networks for lithofacies classification from conventional well logs: An integrated approach," *Petroleum Res.*, vol. 9, no. 3, pp. 393–408, Sep. 2024.

[24] J. Wang, J. Li, K. Li, Z. Li, Y. Kang, J. Chang, and W. Lv, "Borehole lithology modelling with scarce labels by deep transductive learning," *Comput. Geosci.*, vol. 192, Oct. 2024, Art. no. 105706.

[25] D. Datta, G. Singh, S. K. Singh, M. Jenamani, and A. Routray, "Application of multivariate change detection in automated lithofacies classification from well-log data in a nonstationary subsurface," *J. Appl. Geophys.*, vol. 215, Aug. 2023, Art. no. 105094.

[26] S. Tewari and U. D. Dwivedi, "A novel automatic detection and diagnosis module for quantitative lithofacies modeling," in *Proc. Abu Dhabi Int. Petroleum Exhib. Conf.*, Nov. 2018, Paper D012S122R001.

[27] A. Prasad and S. Chandra, "Defending ARP spoofing-based MitM attack using machine learning and device profiling," in *Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS)*, Nov. 2022, pp. 978–982.

[28] S. Tewari, "Assessment of data-driven ensemble methods for conserving wellbore stability in deviated wells," in *Proc. SPE Annu. Tech. Conf. Exhib.*, Sep. 2019, Paper D023S103R018.

[29] S. Tewari and U. D. Dwivedi, "A real-world investigation of TwinSVM for the classification of petroleum drilling data," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Kolkata, India, Jun. 2019, pp. 90–95.

[30] R. Zhong, R. L. Johnson, and C. Zhongwei, "Using machine learning methods to identify coal pay zones from drilling and logging-while-drilling (LWD) data," *SPE J.*, vol. 25, no. 3, pp. 1241–1258, 2020.

[31] J. Jiang, F. Luo, H. Zhang, S. Yang, and Y. Zhang, "Adaptive multiexpert learning for lithology recognition," *SPE J.*, vol. 27, no. 6, pp. 3802–3813, Dec. 2022.

[32] D. Datta, G. Singh, A. Routray, W. K. Mohanty, and R. Mahadik, "Automatic classification of lithofacies with highly imbalanced dataset using multistage SVM classifier," in *Proc. 47th Annu. Conf. IEEE Ind. Electron. (IECON)*, Toronto, ON, Canada, Oct. 2021, pp. 1–6.

[33] R. Qalandari, R. Zhong, C. Salehi, N. Chand, R. L. Johnson, G. Vazquez, J. Mclean-Hodgson, and J. Zimmerman, "Estimation of rock permeability scores using machine learning methods," in *Proc. SPE Asia–Pacific Oil Gas Conf. Exhib.*, Adelaide, SA, Australia, Oct. 2022, Paper D021S008R001.

[34] S. Jiang, P. Sun, F. Lyu, S. Zhu, R. Zhou, B. Li, T. He, Y. Lin, Y. Gao, W. Song, and H. Xu, ''Machine learning (ML) for fluvial lithofacies identification from well logs: A hybrid classification model integrating lithofacies characteristics, logging data distributions, and ML models applicability,'' *Geoenergy Sci. Eng.*, vol. 233, Feb. 2024, Art. no. 212587.

[35] Z. Xiang, P. Li, M. Chadli, and W. Zou, ''Fuzzy optimal control for a class of discrete-time switched nonlinear systems,'' *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 4, pp. 2297–2306, Apr. 2024.

[36] Z. Xiang, P. Li, W. Zou, and C. K. Ahn, ''Data-based optimal switching and control with admissibility guaranteed *Q*-learning,'' *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 5, 2024, doi: 10.1109/TNNLS.2024.3405739.

[37] I. Tomek, ''Two modifications of CNN,'' *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[39] H. He, Y. Bai, E. A. Garcia, and S. Li, ''ADASYN: Adaptive synthetic sampling approach for imbalanced learning,'' in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[40] S. Barua, M. M. Islam, X. Yao, and K. Murase, ''MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, Feb. 2014.

[41] Z. Wu, W. Lin, and Y. Ji, ''An integrated ensemble learning model for imbalanced fault diagnostics and prognostics,'' *IEEE Access*, vol. 6, pp. 8394–8402, 2018.

[42] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2013.

[43] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, ''Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data,'' in *Proc. IEEE Int. Conf. Online Anal. Comput. Sci. (ICOACS)*, May 2016, pp. 225–228.

[44] C. L. Castro and A. P. Braga, ''Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 888–899, Jun. 2013.

[45] K.-H. Huang and H.-T. Lin, ''Cost-sensitive label embedding for multi-label classification,'' *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1725–1746, Oct. 2017.

[46] D. Díaz-Vico, A. R. Figueiras-Vidal, and J. R. Dorronsoro, ''Deep MLPs for imbalanced classification,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–7.

[47] B. Krawczyk, ''Learning from imbalanced data: Open challenges and future directions,'' *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[48] C. J. Peng, K. L. Lee, and G. M. Ingersoll, ''An introduction to logistic regression analysis and reporting,'' *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, 2002.

[49] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsworth, CA, USA: Belmont, 1984.

[50] N. Cristianini and E. Ricci, ''Support vector machines,'' in *Encyclopedia of Algorithms*, M. Y. Kao, Ed., Boston, MA, USA: Springer, 2008.

[51] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, ''Neighbourhood components analysis,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2005, pp. 513–520.

[52] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, ''Tackling the poor assumptions of naive Bayes text classifiers,'' in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, vol. 3, 2003, pp. 616–623.

[53] A. Prasad and S. Chandra, ''VMFCVD: An optimized framework to combat volumetric DDoS attacks using machine learning,'' *Arabian J. Sci. Eng.*, vol. 47, no. 8, pp. 9965–9983, Jan. 2022.

[54] A. Prasad and S. Chandra, ''BotDefender: A collaborative defense framework against botnet attacks using network traffic analysis and machine learning,'' *Arabian J. Sci. Eng.*, vol. 49, no. 3, pp. 3313–3329, Jun. 2023.

[55] S. Wang, L. L. Minku, and X. Yao, ''Resampling-based ensemble methods for online class imbalance learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1356–1368, May 2015.

[56] L. Breiman, ''Bagging predictors,'' *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[57] M. Skurichina and R. P. W. Duin, ''Bagging, boosting and the random subspace method for linear classifiers,'' *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, Jun. 2002.

[58] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[59] Y. Shang, ''Subgraph robustness of complex networks under attacks,'' *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 4, pp. 821–832, Apr. 2019, doi: 10.1109/TSMC.2017.2733545.

[60] P. Geurts, D. Ernst, and L. Wehenkel, ''Extremely randomized trees,'' *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[61] J. H. Friedman, ''Greedy function approximation: A gradient boosting machine,'' *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[62] M. B. Diaz, K. Y. Kim, T. H. Kang, and H. S. Shin, ''Drilling data from an enhanced geothermal project and its pre-processing for ROP forecasting improvement,'' *Geothermics*, vol. 72, pp. 348–357, Mar. 2017.

[63] M. U. A. Bromba and H. Ziegler, ''Application hints for Savitzky–Golay digital smoothing filters,'' *Anal. Chem.*, vol. 53, no. 11, pp. 1583–1586, Sep. 1981.

[64] A. Prasad and S. Chandra, ''PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning,'' *Comput. Secur.*, vol. 136, Jan. 2024, Art. no. 103545.

[65] X. Cui, Y. Li, J. Fan, and T. Wang, ''A novel filter feature selection algorithm based on relief,'' *Appl. Intell.*, vol. 52, no. 5, pp. 5063–5081, Mar. 2022.

[66] K. Baba, L. Bahi, and L. Ouadif, ''Enhancing geophysical signals through the use of Savitzky–Golay filtering method,'' *Geofísica Internacional*, vol. 53, no. 4, pp. 399–409, Oct. 2014.

[67] A. Prasad, S. Chandra, I. Atoum, N. Ahmad, and Y. Alqahhas, ''A collaborative prediction approach to defend against amplified reflection and exploitation attacks,'' *Electron. Res. Arch.*, vol. 31, no. 10, pp. 6045–6070, 2023.

[68] I. G. Roy, ''Optimal Savitzky–Golay derivative filter with geophysical applications: An example of self-potential data,'' *Geophys. Prospecting*, vol. 68, no. 3, pp. 1365–2478, 2019.

[69] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, ''Relief-based feature selection: Introduction and review,'' *J. Biomed. Informat.*, vol. 85, pp. 189–203, Sep. 2018.

[70] D. Chicco and G. Jurman, ''The advantages of the Matthews correlation coefficient (MCC) over F_1 score and accuracy in binary classification evaluation,'' *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020.

[71] S. Boughorbel, F. Jarray, and M. El-Anbari, ''Optimal classifier for imbalanced data using Matthews correlation coefficient metric,'' *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177678.

[72] W.-J. Lin and J. J. Chen, ''Class-imbalanced classifiers for high-dimensional data,'' *Briefings Bioinf.*, vol. 14, no. 1, pp. 13–26, Jan. 2013.

[73] J. Ri and H. Kim, ''G-mean based extreme learning machine for imbalance learning,'' *Digit. Signal Process.*, vol. 98, Mar. 2020, Art. no. 102637.

SAURABH TEWARI received the Ph.D. degree from the Rajiv Gandhi Institute of Petroleum Technology, an institute of national importance, in 2021. Currently, he is an Assistant Professor with the Department of CEA, GLA University, Mathura, India. He has published several high-quality SCI peer-reviewed research works in international journals, attended reputed international conferences, published book chapters, and achieved Indian patent grants along with software copyright. He is researching several machine learning/deep learning-based backend/frontend development projects and the IoT-based low-power chipset edge technology to deploy standalone AI-related solutions. His research interests include artificial intelligence, machine learning model design, testing and deployment, deep learning, transfer learning, and generative AI.

ARVIND PRASAD received the Ph.D. degree in computer science from Babasaheb Bhimrao Ambedkar University (a Central University), Lucknow, with a focus on cybersecurity and machine learning. He is currently an Assistant Professor with the Department of CEA, GLA University, Mathura. Prior to this role, he contributed significantly to academia during his tenure as a Lecturer at King Saud University, Riyadh, from 2010 to 2019. His work is dedicated to developing innovative solutions addressing the dynamic challenges of securing cyberspace. His research interests include cybersecurity, reverse engineering, malware analysis, network traffic analysis, and machine learning.

HARSH PATEL holds a B.Tech. degree in computer science engineering with a data science specialization from the Gyan Ganga Institute of Technology and Sciences in August 2024. His work area is data analytics, machine learning, deep learning, etc. He earned intern experience from the multinational Hexagon company in data analytics, dealing with various aspects of data-driven operations. After completing his B.Tech., he recently joined Humonics Global Private Limited as a Data Scientist. At the undergraduate level, he worked on several research-related projects in various collaborations and research group. He has also won several awards at the national level for his exceptional coding skills in hackathon competitions.

MUEEN UDDIN (Senior Member, IEEE) received the Ph.D. degree from Universiti Teknologi Malaysia (UTM), in 2013. He is currently an Associate Professor of cybersecurity and data sciences with the University of Doha for Science and Technology, Qatar. He has published over 170 international journals and conference papers in highly reputed journals with a cumulative impact factor of over 300. His research interests include blockchain, cybersecurity, the IoT, network security, and cloud computing.

TAHER AL-SHEHARI received the B.Sc. degree in computer science from King Khalid University, Saudi Arabia, in 2007, and the M.S. degree in computer science from the King Fahd University of Petroleum and Minerals (KFUPM), in 2014. From 2011 to 2014, he was a Research Assistant at KFUPM. Since 2015, he has been a Senior Lecturer and a Researcher at King Saud University. He is the author of several papers that are published in prestige journals. His research interests include information security and privacy, insider threat detection and prevention systems, machine learning models, and data analysis. His awards and honors include an Honor Award from King Khalid University's Rector, and a Best Designed Curriculum Award from CFY's Dean, KSU.

NASSER A. ALSADHAN is currently with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He has been instrumental in advancing research in various areas of technology. His work has spanned a range of topics, including prediction accuracy, affective states, and big data analytics. His expertise in convolutional neural networks and language models has led to innovative developments in understanding emotional words and personality traits through text analysis.

Dr. Alsadhan is a Distinguished Member of the IEEE Community, known for his significant contributions to the field of computer and information sciences.

Open Access funding provided by 'Qatar National Library' within the CRUI CARE Agreement