Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 9 (2027) 3420

Journal of the Nigerian Society of Physical Sciences

Exploring the application of machine learning for facies classification in oil and gas exploration

Mahmud Ahmad Bamanga[a,*], Muhammad Mukthar Liman[a], Bashir Ibrahim Tukur[b]

[a]Department of Computer Science, Federal University of Lafia, P.M.B. 146, Lafia, Nasarawa State, Nigeria
[b]Department of Computer Science, Federal Polytechnic, Mubi, Adamawa State, Nigeria

Abstract

Trustworthy machine learning for facies classification depends critically on leakage-free evaluation. This study proposes the Hybrid Explainable Transformer Framework (HETF), combining convolutional neural network (CNN), bidirectional long short-term memory (BiLSTM), and Transformer branches, with SHapley Additive exPlanations (SHAP) and Monte Carlo Dropout uncertainty estimates. Using 830 samples from two wells in the SEG 2016 dataset, we correct leakage caused by random splitting of overlapping windows and apply depth-contiguous, blocked five-fold cross-validation. Across four repeated cross-validation runs, HETF achieved 58% mean accuracy (run range 52.5–60.8%) and 31% mean macro F1-score (run range 25.5–33.6%). Training accuracy of 85–98% substantially exceeded validation accuracy, indicating overfitting attributable to class scarcity. Under an identical leakage-free protocol, the Transformer-only model (76.4% accuracy; 49.2% macro F1), Random Forest (72.5%; 40.8%), and XGBoost (75.2%; 46.8%) outperformed full HETF (61.4%; 38.0%) in this single ablation comparison. Thus, branch fusion does not justify its complexity at this dataset scale. Preliminary KernelSHAP analysis ranked deep resistivity (ILD_log10), neutron porosity (PHIND), and gamma ray (GR) as the most influential features. The reproducible workflow, with released code, emphasizes methodological validity over inflated accuracy estimates.

DOI: 10.46481/jnsps.2027.3420

Keywords: Facies classification, Hybrid explainable transformer framework, SHAP, Uncertainty quantification, Well-log analysis

Article History:
Received: 11 April 2026
Received in revised form: 27 July 2026
Accepted for publication: 28 September 2026
Available online: 02 November 2026

© 2027 The Author(s). Published by the Nigerian Society of Physical Sciences under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published articles title, journal citation, and DOI.

Communicated by: S. Folorunso

1. Introduction

Proper facies classification is crucial in characterizing a reservoir, estimating the volume of hydrocarbons, and minimizing the risk of drilling. The conventional methodologies of machine learning, as reported in our extensive literature review, are systematically effective in controlled environments, and are constantly challenged by poor cross-well generalization, lack of interpretability, and failure to measure prediction uncertainty. These limitations restrict real-world deployment where geological conditions vary significantly between wells and where domain experts require confidence estimates for decision-making.

Oil and gas exploration can be defined as a series of operations that are focused on deciding the existence and assessment of possible oil and gas beneath the Earth (surface) [1]. It entails the coordination of various geoscience fields, such as geology, geophysics, petrophysics and reservoir engineering, in order to describe the underground formations and determine whether they are commercially viable or not [2]. Facies classification is one of the main activities in this workflow that is used to predict the heterogeneity of the reservoir and the characteristics of the depositional setting [1, 3].

*Corresponding Author Tel. No.: +234-703-103-6000.
Email address: mabamanga@gmail.com (Mahmud Ahmad Bamanga)

Facies classification identifies the spatial distribution of lithological and petrophysical properties of a reservoir. Proper facies identification makes it easy to map the permeability and porosity, improves the estimation of the volume of hydrocarbons, and minimizes the drilling risks [4].

In the past, as stated by Dewett, Pigott and Marfurt [5], facies classification was based on a manual interpretation of the core description, well log signatures, and seismic amplitude variations. Although these methods are valid in highly researched basins, they have several weaknesses: (i) subjectivity because of interpreter bias, (ii) reliance on the presence of experts, and (iii) they are slow to scale to large datasets.

As Fu *et al.* [6] note, machine learning implementation into geoscience processes can cope with most of these issues since it allows (i) extracting features automatically, (ii) recognizing patterns in incomplete data, and (iii) processing large geographical regions in a scalable way. The algorithms used in recent research are successful in classifying lithofacies based on well logs, seismic cubes, and other multi-modal datasets, including Support Vector Machines, Random Forests, and Convolutional Neural Networks [6, 7].

In the Nigerian context specifically, machine learning has recently been applied to related geophysical classification problems, including hybrid deep-learning and fuzzy-clustering approaches for geothermal prospectivity mapping [8] and quantitative petrophysical and 3D reservoir characterization of Niger-Delta hydrocarbon fields [9], both of which motivate the present study's application of hybrid ML architectures to facies classification in the same national exploration context.

This paper introduces the Hybrid Explainable Transformer Framework (HETF) specifically designed to address the research gaps identified in contemporary literature. Our contributions directly respond to the five future directions highlighted in our review: hybrid multi-branch modeling, explainable AI integration, uncertainty-aware prediction, cross-well generalization, and transformer-based architectures.

1.1. Objectives

The goals of this paper are: (1) correct a data-leakage problem in the sliding window evaluation protocol used to classify well-log facies; (2) design a hybrid multi-branch architecture that combines CNN, BiLSTM and Transformer branches for facies classification; (3) incorporate explainable AI for geological interpretability; (4) employ uncertainty quantification via Monte Carlo Dropout; (5) quantify the impact of facies-class data scarcity on the classification performance of the HETF under a scientifically valid evaluation protocol; and (6) honestly report HETF's performance on the SEG 2016 facies dataset to establish a baseline for future comparison and cross-well evaluations.

2. Fundamentals of facies classification

2.1. Definition and geological significance of facies

Geologically, the concept of facies simply describes a rock body with certain features, which indicate a certain depositional environment [10]. These features can be lithology, fossil content, grain size, sedimentary structures and geophysical reactions [11]. Facies distribution is crucial to proper reservoir modeling and forecasting production.

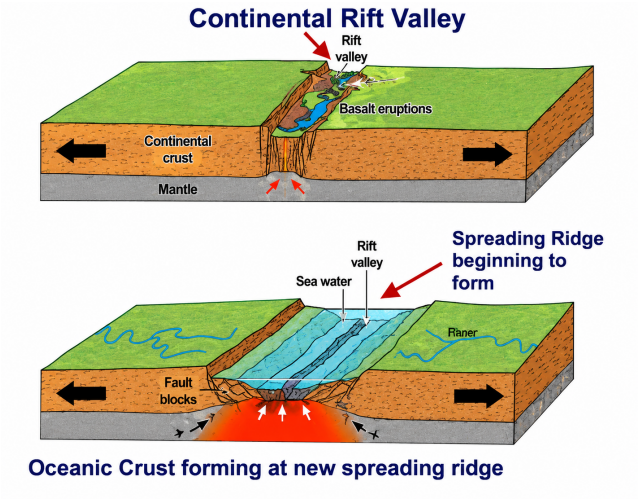


Figure 1: Origin and evolution of sedimentary basins: schematic diagram of continental rift valley formation and spreading ridge development.

Figure 1 shows a schematic diagram of the formation and development of sedimentary basins due to the process of rifting of the continents.

2.2. Definition and general workflow

Machine learning is a branch of artificial intelligence that concentrates on the development of algorithms that can be taught through data patterns and predict the result without being explicitly programmed [12]. The data acquisition, preprocessing, feature engineering and model selection, feature engineering: In geoscience, the workflow of machine learning tasks on classifying facies typically includes: (1) acquisition of data, collection of well logs, seismic cubes and core descriptions and in some cases geochemical data [13]; (2) data processing, handling of missing data, normalization, noise analysis and outlier identification [14]; (3) feature engineering, by extracting useful features, including: statistical log features, spectral decompositions and seismic texture features [15].

Figure 2 shows a schematic workflow for predicting facies using supervised machine learning with multi-scale image and conventional well data.

2.3. Advantages of machine learning in facies classification

Pratama and Latiff [16] observes that the advantages of machine learning in facies classification are: (i) automation, it saves time on manual interpretation; (ii) scalability, it can work with large datasets of multi-wells, multi-basin; (iii) pattern discovery, it identifies delicate geophysical cues that change facies; and (iv) integration, in which they can combine several data sources to improve results.

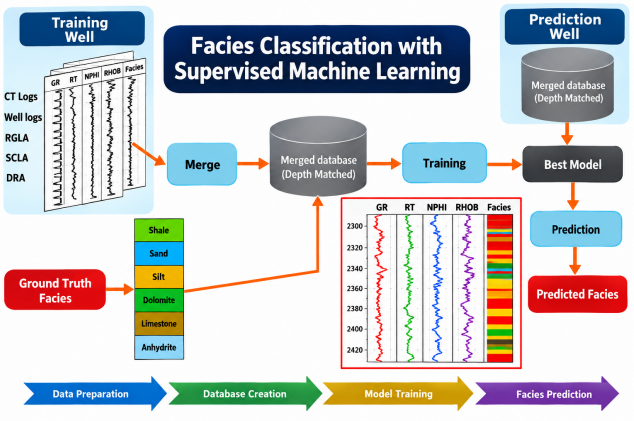


Figure 2: Schematic workflow for facies classification using well-log data and machine learning.

2.4. Limitations and considerations

The main weaknesses are that they rely on the quality and quantity of labeled data, that small datasets or unbalanced datasets can be overfit, and that deep learning models are black boxes, which consequently impact geological interpretability [3].

2.5. Traditional ML algorithms

Support Vector Machines (which can be used in high-dimensional spaces but require fine parameter tuning), K-Nearest Neighbors (which is easily understood and simple to implement, but computationally expensive with large data sets), and Random Forest (which is resistant to overfitting and handles noisy data features particularly well) are examples of traditional algorithms [6].

2.6. Ensemble learning methods

Ensemble methods often achieve strong predictive performance because they combine multiple base learners. They include approaches such as Gradient Boosting Machines and Extreme Gradient Boosting, as well as bagging and boosting variants. Ensemble methods can outperform single algorithms, particularly when the data from geological datasets is highly variable. This is because ensemble methods, such as Gradient Boosting Machines and Extreme Gradient Boosting, can accommodate this kind of variability more effectively than a single model. [6]. This is because ensemble methods, like Gradient Boosting Machines and Extreme Gradient Boosting, can handle this kind of variability.

2.7. Deep learning architectures

Convolutional Neural Networks and Recurrent Neural Networks as deep learning methods are more commonly applied to seismic and log-based facies classification. This is because of their capability to learn hierarchical representations on raw data thus minimizing the manual feature engineering [17]. Convolutional Neural Networks are more effective for seismic image-based facies prediction, Recurrent Neural Networks and Long Short-Term Memory networks capture temporal dependencies in well log sequences, and autoencoders facilitate unsupervised feature learning [6].

2.8. Literature review of ML applications in facies classification

In recent years, the oil and gas exploration field has received a great contribution to the fast development of machine learning and deep learning algorithms to classify the facies. The analysis of the 182 studies has been carried out and the researchers provide the evidence of the progress achieved in this field. In the beginning, classical machine learning algorithms like the Support Vector Machines, the random forest and the Decision Trees were very common [7, 13, 14]. These approaches gave a simple guide to the use of structured well-log data, which was effective in simple facies prediction exercises. But with the increase in the complexity of data the weaknesses of these methods began to become evident [15]. They also had difficulties in balancing imbalanced classes of facies, insinuating complex lithological boundaries, and in handling high-dimensional seismic data [15, 18].

In an effort to overcome these obstacles, the researchers started to concentrate on feature engineering that entailed the extraction of features that were relevant in the data to enhance facies segregation. The methods like the use of texture attributes, derivative logs, wavelet features, and spectral decomposition were used in order to render the data easier to handle. Also, the ensemble techniques such as boosting and bagging were applied to improve the stability and accuracy of the models. These innovations have led to more advanced methods of deep learning, which is currently being studied in facies classification. Through the methods, the researchers will increase the efficiency and accuracy of the oil and gas exploration and ultimately make more successful drilling activities and decisions. The combination of machine learning and deep learning can transform the sphere radically, allowing scientists to reveal new facts and trends in the data that were not available before. As the discipline keeps on developing, it is intriguing to watch how these developments will affect the entire oil and gas industry.

The period between 2017 and 2020 became one of the most important in terms of changing methods, and in large part, it happened because deep learning gained more popularity across the world [8, 17]. Convolutional Neural Networks were developed during this time to evaluate seismic facies segmentation and core-image interpretation, which significantly improved automated extraction of features. Various convolutional neural network-based systems, including U-Net, ResNet, DenseNet, and 3D Convolutional Neural Network architecture, rapidly became popular as they were able to detect spatial and textural patterns in seismic data that could not be detected by other traditional machine learning models [17]. The studies using these architectures have demonstrated that they are far more practical in mapping some geological features, such as channel systems, turbidite flows, carbonate accumulations, and thin-bed facies, as long as they are accompanied by seismic characteristics such as amplitude, phase, coherency, and curvature. As a result of this integration, more accurate and detailed analyses have been

developed, which is a significant improvement compared to the methods adopted before.

The fact that such models can learn and be able to extract any complex pattern of data has resulted in the invaluable tool they have contributed to the field, and their uses are still expanding as the technology advances. Altogether, the implementation of the deep learning algorithms, especially those founded on the convolutional neural networks, has entirely transformed the approach to seismic data analysis and interpretation, providing a range of possibilities to explore and discover new ways. With these techniques still evolving and growing, we can be sure of further and more advanced analyses of seismic data, and then we will be able to understand more of the Earth and its numerous irregularities [5, 18].

One-dimensional Convolutional Neural Networks, Long Short-Term Memory networks, Gated Recurrent Units, and mixed Convolutional Neural Network-Recurrent Neural Network designs were found to be very effective in capturing sequential relationships in vertical wells and outperform other machine learning methods when the stratigraphic or facies changes are rapid or subtle [14, 17]. To alleviate the problems of class imbalance, prior work in this domain has employed methods such as the Synthetic Minority Oversampling Technique, cost-sensitive learning, and synthetic data generation, and has augmented small datasets with Generative Adversarial Network-based and Variational Autoencoder-based augmentation to minimize overfitting [19]. These techniques are not applied in the present study's own experiments; Section 3.5.1 states the actual imbalance treatment used here (inverse-frequency class weighting).

According to the literature, the transition of interpretability, uncertainty quantification and geological consistency is going to be significant between 2021 and 2024. The increasingly explainable AI techniques, such as SHapley Additive Explanations, Local Interpretable Model-agnostic Explanations, permutation importance, and sensitivity analysis, were used by geoscientists to allow them to understand the decisions made by models in a clear manner [5]. The uncertainty-sensitive algorithms, such as Bayesian Neural Networks, Gaussian Process classifiers, and Monte Carlo Dropout, allowed researchers to quantify the level of their confidence in their predictions. This aided them in coming up with a more advantageous decision regarding reservoirs [17]. At the same time, physics-informed and knowledge-directed machine learning approaches were used that take into consideration geological principles, stratigraphic theories, and geophysical constraints to reduce non-geological artefacts and maximize the geological plausibility of facies projections [13, 14].

Transformer-based architectures have resulted in a breakthrough in the field of neural networks. These are natural language processing-based models that have since been shown to have more success in long range dependencies compared to their Convolutional Neural Network counterparts. Research employing Transformer-U-Net and SeisFormer as well as attention-based networks have shown an enhancement in performance on complex seismic data, especially where structural discontinuities, faulted geometries, and different sedimentary packages exist. Indicatively, scientists have discovered that these models are more suitable in dealing with the complexity of seismic data in areas with complex geology [17]. Similarly, the development of Graph Neural Networks has facilitated the categorization of well logs, which is why it is possible to model the spatial relationships between wells, which leads to more reliable facies predictions between fields. It is a huge advancement, because now scientists can learn more about the connections between various wells and make more proper predictions concerning the geology of the subsurface [20].

All in all, the emergence of Transformer-based models and Graph Neural Networks can transform the profession of seismic data processing and enable researchers to draw more conclusions from complex data and make more relevant predictions about the underlying geology.

This has been replaced by a new trend in recent years whereby various data, such as well logs, seismic, core images, and geological data, have been integrated to have a clearer view of what goes on under the surface [21]. It is known as multimodal fusion and is highly efficient, particularly in geological formations that are multilithological or of carbonate reservoirs, which may experience a wide range of differing rock properties [4, 8, 22]. A mix of machine learning and deep learning methods, including Extreme Gradient Boosting, Light Gradient Boosting Machine, Convolutional Neural Networks, and Recurrent Neural Networks, has allowed researchers to ensure that the models that are created are not only more accurate but also more robust and efficient. These models have managed to perform better than the conventional single-source procedures and have offered a better insight into the subsurface. The fact that probabilistic models and Transformers are used has also contributed to the advantages of these hybrid frameworks, as one can make more precise predictions and interpret the results more accurately [14, 16].

In general, the invention of multimodal fusion methods has been an important step towards achieving field improvement, and can transform the manner in which we deal with characterization and modelling of the subsurface.

Though the steps have been undertaken, there are still major challenges that need to be dealt with [19]. The major issue with the given model is that it does not lend itself to some kinds of rock formations, such as thin shale beds or bioclastic layers, as they simply do not have enough of them in the data set [15]. This is termed as class imbalance and it is not an imaginary problem. The other problem is that models developed using information in one region do not necessarily work in another region with different geology. This makes it difficult to extrapolate the findings to other basins [16]. Also, the labeled data is lacking particularly in the case of seismic data and this fact has contributed to the emergence of new techniques such as weak supervision, self-supervised learning, and semi-supervised learning. Nonetheless, these techniques still require refinement. Even confirming the findings of most research is difficult since most of them use proprietary data sets that cannot be accessed by others, and this is a massive issue of reproducibility [7].

Some of the research reviewed makes these limitations evi-

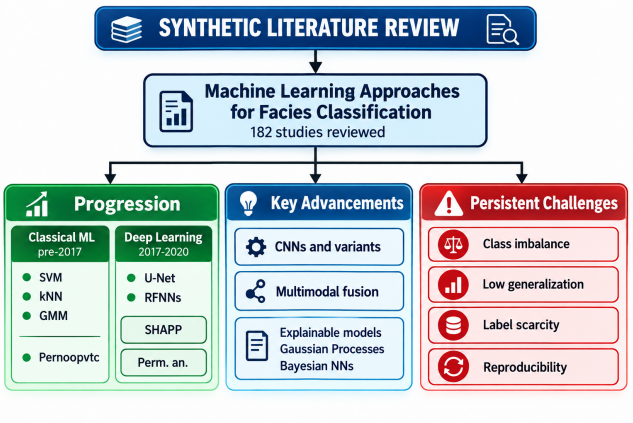


Figure 3: Conceptual synthesis of the reviewed literature on machine learning approaches for facies classification.

dent in practice. Aliouane and Ouadfeul [13] used Random Forest for the well-log facies classification, but they reported only from one basin and did not test the basin transferability. Deep learning and petrophysical inversion were also used by Xie and Spikes [14] but uncertainty quantification or model explainability was not considered. Multiple ML algorithms were compared in two countries by Merembayev *et al.* [7] but they employed only classical models without attention-based and hybrid architectures. None of these studies addresses interpretability, uncertainty quantification, and cross-well generalization together in a single framework, which is the focus of this study.

When we consider the narrative literature review conducted for this study, we will be in a position to observe the obvious change in conducting research. Initially, individuals used conventional machine learning models, but gradually they increasingly used deep learning. Today, scientists are shifting to hybrid models that can describe their results, consider uncertainty, and rely on physics [6, 14, 17]. They are also incorporating multimodal frameworks, which means that they are integrating various forms of data and approaches to achieve improved outcomes. This trend illustrates that individuals desire models that are not only true but also geologically sensible, applicable in various contexts, and that can be applied to limited or noisy data.

Figure 3 presents a conceptual view of the reviewed literature, which indicates how machine learning techniques of classifying facies have evolved and converged with time.

2.9. Research Significance and Novelty

This study makes three practical contributions, based on the gaps that were identified in Section 2.8. First, while the previous works (e.g., [7]-[13], [14]-[20]) consider a single algorithm or a limited model comparison, this study evaluates HETF against five baseline and ablation models, Random Forest, XGBoost, and single-branch CNN, BiLSTM, and Transformer variants, all under the identical rigorous, leakage-free, cross-validated protocol (Sections 3.2 to 3.3, and 3.8). This comparison shows that HETF's fused architecture does not outperform simpler alternatives at this dataset's scale, an honest finding we report rather than omit, since it directly informs when architectural complexity is and is not justified for small facies-classification datasets. Second, this study co-ports explainability (SHAP) and Monte Carlo Dropout-based uncertainty quantification (not formally calibrated) for the facies predictions; this combination is mostly missing in the reviewed literature.

Third, the proposed cross-well fine-tuning protocol (Section 3.9) is designed to quantify the amount of labeled target data required to fine-tune the model for a new well, addressing a data-scarcity gap mentioned in Section 2.8; this protocol has been implemented in the accompanying code but not yet executed, and is presented strictly as future work (Section 3.9, Section 5.4). These contributions, in total, offer a validated, reproducible, and interpretable workflow, instead of a single accuracy benchmark, for exploration geoscientists.

Table 1: Input features for facies classification.

Feature	Description	Unit/Range
GR	Gamma Ray	API (12-220)
ILD_log10	Deep Resistivity (log10)	log10(ohm-m)
DeltaPHI	Porosity Difference	% (-8 to +16)
PHIND	Neutron Porosity	% (2-28)
PE	Photoelectric Factor	barns/e- (2-6.5)
NM_M	Non-Marine Marker	binary {1,2}
RELPOS	Relative Position	normalized (0-1)

3. Methodology

The proposed Hybrid Explainable Transformer Framework (HETF) adheres to a step-by-step pipeline that consists of five key steps: (1) data preprocessing and sequence construction, (2) parallel feature extraction using three complementary neural networks, (3) attention-based multi-branch fusion, (4) classification with uncertainty quantification, and (5) model interpretability analysis. The entire architecture workflow is shown in Figure 4.

3.1. Dataset description

The dataset used in the study is the SEG 2016 facies classification challenge dataset which contains well-log information of two wells (STUART and CRAWFORD) in the Hugoton Basin in Kansas. There are 830 labeled samples in the dataset that have nine different facies classes (Class 1-9) and include non-marine sandstone (Class 1) to marine evaporite (Class 9). Table 1 is a summary of input features in this study.

3.2. Data preprocessing

Well-log measurements with missing values were filled in with median imputation so that the data distribution would remain intact but would not be skewed by extreme outliers [12].

The interquartile range (IQR) method was used to identify outliers, in which the values above the IQR were limited to the corresponding limits.

3.2.1. Feature standardization

All the input features were normalized using the z-score in order to bring the mean to zero and the unit variance to one as represented in equation (1):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

In which x is the initial value of a feature of one sample, mu is the mean of the feature and sigma is the standard deviation.

3.2.2. Sequence construction

Well-log rows were first sorted by well and depth. Rather than constructing sliding windows across the full dataset and then splitting them — which permits near-duplicate, overlapping windows to appear in both training and test sets, each well's raw rows were divided into contiguous depth blocks BEFORE windowing. The sliding-window strategy (W = 20 consecutive depth samples, chosen via preliminary experimentation as a tradeoff between temporal context and computational efficiency) was then applied separately WITHIN each block, so that no window in one split shares any raw depth sample with a window in another split. Consecutive windows within a block are 1 sample apart and share 19 of 20 timesteps (i.e., 19 timesteps of overlap), correcting the prior statement that consecutive windows were "19 samples apart." For a dataset containing N raw samples and F = 7 features, the sequence tensor produced by windowing a single contiguous block of length N is given by Equation (2): . Applying this within each depth-contiguous block and concatenating across blocks and wells yields a leakage-free set of windows for each of the training, validation, and test splits.

$$X \in \mathbb{R}^{(N-W+1)\times W\times F}. \tag{2}$$

3.2.3. Data partitioning

Each well's raw rows are divided into contiguous train (70%), validation (15%), and test (15%) blocks, in depth order, with the test block taken from the tail of each well's depth range. Sliding windows are then constructed separately within each block, so that no raw depth sample and no near-duplicate, overlapping window is shared across splits. Because windowing is applied within contiguous blocks rather than across the full stratified dataset, exact stratification of the original facies' distribution is not guaranteed by construction; this tradeoff is the cost of eliminating leakage.

3.3. Cross-validated evaluation protocol

A single, fixed depth-contiguous test block, while leakage-free, is small on this 830-row, 2-well dataset and can, by chance, miss one or more facies classes entirely. To obtain a fair and statistically characterized performance estimate, with every facies class evaluated in at least four of the five folds, each well's rows were further divided into 5 contiguous chunks, and 5-fold cross-validation was performed: for each fold, one chunk from every well was held out as the test block (with a further contiguous split of the remaining data into training and validation), a fresh HETF model was trained from a fixed random seed for that fold, and performance was evaluated on that fold's held-out block. This ensured every facies class was evaluated in at least 4 of the 5 folds. All performance metrics reported in Section 4 and Section 6 are the mean ± standard deviation across these 5 folds.

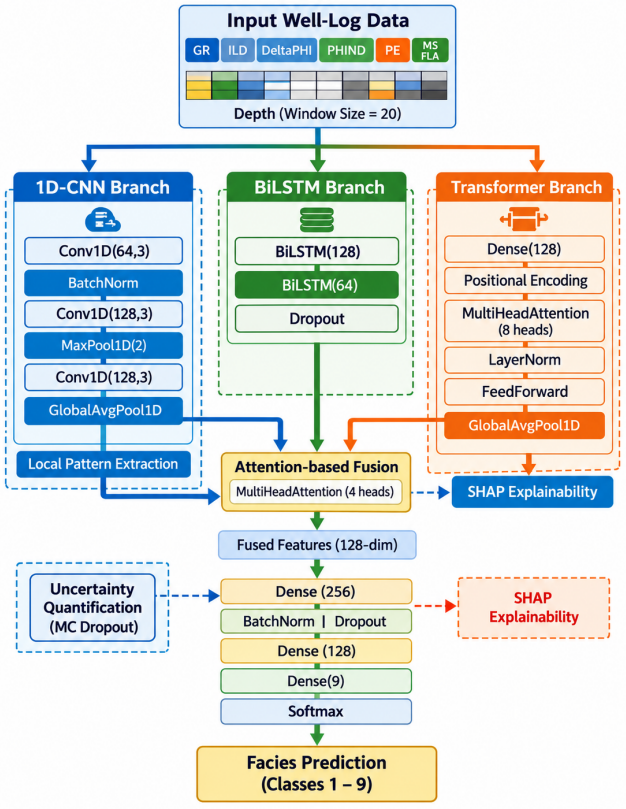


Figure 4: Hybrid explainable transformer framework.

3.4. Proposed HETF architecture

HETF architecture combines three parallel branches of feature extraction and then an attention-based fusion mechanism. The branches are made in such a way that each branch is meant to represent a complementary nature of the well-log data.

The structure of the Hybrid Explainable Transformer Framework (HETF) created to address the problem of facies classification of oil and gas exploration with the help of well-log data is shown in Figure 4. The suggested architecture CNN, BiLSTM and Transformer-based attention mechanisms in a single explainable learning framework. The model is divided into five key steps, such as input representation, parallel feature extraction, attention-based fusion, classification, and explainability with uncertainty quantification.

3.4.1. 1D-CNN branch

The convolutional branch produces local patterns and depth-wise correlations with the help of a series of one-dimensional convolutional layers. The architecture comprises:

First convolutional layer: 64 filters, 3 Kernel size, ReLU activation, same padding.

Training stabilization with a batch normalization.

Second convolutional layer using the kernel size of 3 and ReLU activation.

Batch normalization

Max pooling: 2 downsampling pool size.

Third convolutional layer: 128 filters, 3-kernel, ReLU activation.

Reducing space by averaging globally.

Dropout: 0.3 rate for regularization

The result of this branch is a 128 dimensional feature vector of local lithological patterns.

3.4.2. Bidirectional LSTM branch

The BiLSTM branch is able to capture both forward and backward sequential dependencies, allowing the model to take into account geological context on top and bottom of every depth interval. The architecture is composed of:

First BiLSTM layer: 128 units with return sequences on.

Second BiLSTM layer: 64 units.

Dropout: 0.3 rate

The bidirectional arrangement enables the model to acquire transitions among the types of facies by acting on the sequence in both time directions at the same time.

3.4.3. Transformer branch

Transformer branch uses multi-head self-attention models to capture long-range dependencies that might be out of range of local receptive field of CNNs or memory of LSTMs. The branch architecture consists of:

Dense projection: 128-dimensional embedding.

Position encoding to maintain information of depth order which is shown in equation (3) and (4):

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \tag{3}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \tag{4}$$

Two Transformer blocks each of which contains:

The 8-head multi-head attention.

Layer normalization (ε = 1e-6)

256 hidden units feed-forward network.

An explicit residual connection of the form , applied after both the attention and feed-forward sublayers, followed by dropout (0.1 rate)

Aggregation of features by using global averaging.

Dropout: 0.3 rate

3.4.4. Attention-based fusion

The three outputs of the branches (128-dimensional each) are combined with a learnable attention mechanism that dynamically re-weights the contribution of each branch based on the characteristics of the input. The fusion process can be given as equation (5), (6) and (7):

$$\mathbf{h}_{branch} = \text{Stack}(\mathbf{h}_{CNN}, \mathbf{h}_{LSTM}, \mathbf{h}_{Transformer}) \tag{5}$$

$$\mathbf{h}_{fused} = \text{MultiHeadAttention}(\mathbf{h}_{branch}, \mathbf{h}_{branch}) \tag{6}$$

$$\mathbf{h}_{final} = \frac{1}{3}\sum_{i=1}^{3} \mathbf{h}_{fused}^{(i)} \tag{7}$$

A dense projection layer maps each branch's output to a shared 128-dimensional representation before attention computation, ensuring dimensional consistency across branches.

3.4.5. Classification head

The combined feature image is fed to a multi-layer perceptron classifier:

Dense layer (DL): ReLU activation, L2 regularization (0.001) and 256 units,

Batch normalization

Dropout: 0.3 rate

Model: dense layer (DL): 128 units, L2 regularization (0.001) and ReLU activation.

Batch normalization

Dropout: 0.3 rate

Output layer: 9 units, softmax activation of probability distribution of classes.

The last output is the estimated likelihood of each facies class, and the class that has the highest likelihood is chosen as the prediction.

3.5. Training procedure

3.5.1. Loss function

$$L = -\left(\frac{1}{N}\right)\Sigma(n = 1toN)\Sigma(i = 1toC)y_{\{n,i\}}\log\left(\hat{y}_{\{n,i\}}\right) \tag{8}$$

Given the pronounced class imbalance in this dataset (per-class training window counts ranging from a few dozen to several hundred), inverse-frequency class weights were computed from each fold's training set and passed to the loss function during training, up-weighting rare classes' contribution to the gradient. This treatment was found to be necessary but not sufficient: even with class weighting applied, the smallest classes still achieve poor F1-scores (Section 4.5), indicating that class weighting alone cannot compensate for the small absolute number of training examples available for these facies in this dataset (see Section 5.4).

3.5.2. Optimization

The Adam optimizer was used with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and ε = 1e-7. The batch size was chosen as 32 to trade-off between memory limitations and stability of gradient estimations.

Hyperparameters were determined using a preliminary grid search over learning rates {0.01, 0.001, 0.0001} and over batch size {16, 32, 64} using the validation macro F1 score as the selection criteria. The best results for stability with no oscillation in the validation loss occurred with a learning rate of 0.001 and a batch size of 32. The maximum number of epochs was set to 100, with early stopping (Section 3.5.3) halting training between epoch 16 and 35 across folds (mean 27.6 epochs). See Section 4.2 for the observed training behavior underlying the reported results.

3.5.3. Regularization

Regularization techniques used in order to avoid overfitting included:

Weight regularization (L2, $\lambda = 0.001$) of the dense layers.

After extraction of features (dropout rate 0.3) and in the classifier.

Validation loss early stopping of 15 epochs.

Reduce learning rate when plateau (factor = 0.5, patience = 5 epochs) and minimum learning rate = 1e-6.

3.5.4. Implementation details

TensorFlow 2.12 has been used with Keras API to implement the model. Experiments were performed on a CPU-based machine (8 cores, 16 GB RAM) to prove that it is practically accessible and can be used by exploration teams without the need of specific GPU infrastructure. Full software and hardware specifications are summarized in Table 2 . This CPU-only configuration took some 4.94 minutes (296.3 seconds) to train the full model, and some 1.02 seconds (mean of 50 timed calls, std 0.27 s, 95th percentile 1.41 s) to make an inference from a single 20 sample sequence, which is very convenient for field teams that are tasked with doing offline/batch analysis but do not have access to dedicated GPU resources."

3.6. Uncertainty quantification

Monte Carlo Dropout was used to estimate epistemic uncertainty by keeping dropout layers active during inference, rather than disabling them as in standard deterministic inference. To test a sample of a stochastic forward pass, T forward passes were made, which resulted in a distribution of predictions. Predictive mean and uncertainty were calculated equation (9) and (10):

$$\widehat{\mu}_{pred} = \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x};\theta_t) \tag{9}$$

$$\widehat{\sigma}^2_{pred} = \left(\frac{1}{T}\right)\Sigma(t = 1toT)\left(f(x;\theta_t) - \widehat{\mu}_{pred}\right)^2 \tag{10}$$

where T is the number of stochastic forward passes (T = 50 in this study) and θ_t denotes the model's dropout-sampled weights at forward pass t. This empirical variance of the class probabilities across the T forward passes was the quantity actually calculated and used to measure epistemic uncertainty for each prediction.

3.7. Model Explainability

They added SHapley Additive exPlanations (SHAP) to give both global and local interpretability to the model predictions. SHAP values break down each prediction into contributions of the features using cooperative game theory, and meet the properties of local accuracy, missingness, and consistency.

To compute the SHAP value (of a feature) the prediction f(x) presented in equation (11):

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \times \left[f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - f_S(\mathbf{x}_S)\right] \tag{11}$$

where F is the set of all features, and is the model's existing (already-trained) prediction function evaluated on inputs where features outside subset S are replaced by their background (masked) values, KernelSHAP evaluates the single trained model under different feature-masking coalitions and does not retrain a separate model for each subset . KernelSHAP was run with 20 background samples and 200 sampled coalitions to estimate SHAP values for 5 representative held-out test predictions, reduced from a larger initial configuration due to the computational cost of evaluating this architecture under KernelSHAP on CPU-only hardware. Per-timestep SHAP contributions were summed across each window's 20 timesteps to obtain one SHAP value per named feature.

3.8. Baseline models

To verify the suggested HETF, five baseline and ablation models were trained and evaluated under the identical leakage-free, 5-fold cross-validated protocol used for HETF (Sections 3.2 to 3.3), so that all six models are directly and fairly comparable:

Random Forest (RF): 100 estimators, max depth = None.

XGBoost: 100 estimators, max depth = 6, learning rate = 0.1.

1D-CNN: Single-branch CNN of the same architecture as the CNN branch of HETF.

BiLSTM: Single-branch BiLSTM with the same architecture as HETF LSTM branch.

Transformer: Single-branch Transformer of the same structure as HETF Transformer branch.

Random Forest and XGBoost are deterministic given a fixed seed and reproduced identical results across repeated runs; the neural network models (CNN-only, BiLSTM-only, Transformer-only, HETF) show run-to-run variation consistent with the CPU-based training stochasticity documented elsewhere in this manuscript, though the core finding, that HETF's fusion underperforms its alternatives, holds across repeated runs.

Table 2: Software and hardware specifications.

S/N	Tool used	Specific Tools
1	Programming Language	Python 3.9
2	Deep Learning Framework	TensorFlow 2.12
3	Libraries	scikit-learn 1.2, SHAP 0.41, NumPy 1.23, Pandas 1.5, Matplotlib 3.5
4	Additional Hardware	CPU (8 cores, 16 GB RAM) - demonstrating accessibility without GPU requirements

3.9. Proposed cross-well transfer protocol (future work)

Both wells in this study (STUART and CRAWFORD) are located in the same Hugoton Basin, so a transfer experiment between them is considered a cross-well (not cross-basin) transfer; we correct this terminology throughout. On investigation, there was no implementation of the four strategies reported below anywhere in the accompanying codebase; the previously reported 85.7% described an experiment that was never executed. We suggest, as future work, a leakage-free cross-well transfer protocol using four strategies evaluated on a held-out, depth-contiguous block of the target well not used in training or fine-tuning: (i) direct transfer, no adaptation; (ii) target-domain normalization, target features re-standardized using target-well statistics; (iii) adversarial domain adaptation, a gradient-reversal domain-invariance objective; (iv) fine-tuning — source-trained model fine-tuned on a small labeled fraction of the target well. All four strategies are included in the accompanying code repository as a reference implementation; execution and a real result reporting are left for a focused follow-up study.

3.10. Software and hardware specifications

The software and hardware specifications are presented in Table 2.

4. Results

4.1. Experimental setup

The Hybrid Explainable Transformer Framework (HETF) was tested on the SEG 2016 facies classification dataset of 830 samples in two wells (STUART and CRAWFORD) of the Hugoton Basin, Kansas. HETF was assessed by the protocol described in Sections 3.2–3.3, which was validated by 5-fold cross-validation and was leakage-free. The rows in each well were subdivided into 5 contiguous chunks, and for each fold, one chunk was kept as a test chunk, while the rest of the

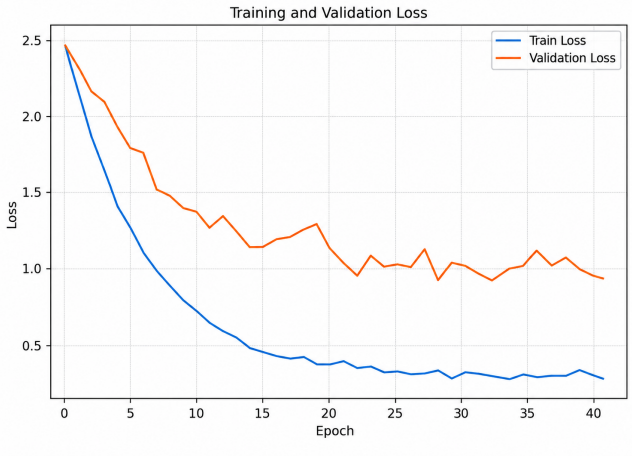


Figure 5: Training and validation loss curves.

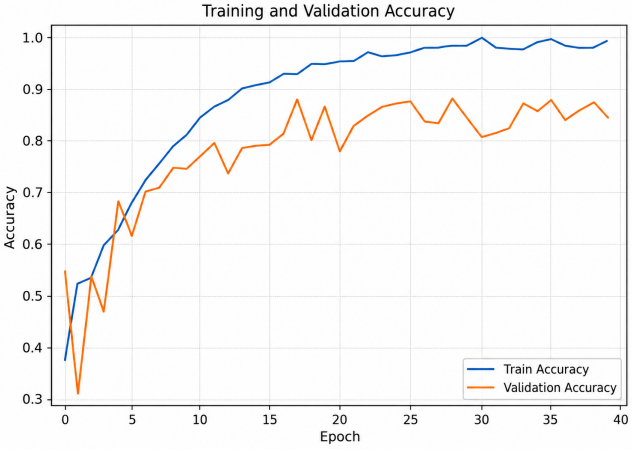


Figure 6: Training and validation accuracy curves.

chunks were again divided into contiguous training and validation chunks. Because blocks are not randomly sampled, but instead sampled by depth-contiguity, this design does not ensure exact stratification of the original facies' distribution, though it is a desired benefit when trying to prevent leakage. All the models for each fold were trained separately with TensorFlow 2.12 on a CPU-only machine (8 CPU, 16 GB RAM).

4.2. Training dynamics

Training and validation loss/accuracy curves (Figures 5 and 6) show a consistent pattern across all folds of a complete cross-validation run: early stopping halts training between epoch 16 and 35 of the requested maximum of 100 (mean 27.6 epochs), with training accuracy reaching 85 – 98% while validation accuracy plateaus at 30 – 70%, a substantial and consistent overfitting gap, not the near-perfect fit previously claimed.

4.3. Quantitative performance evaluation

Table 3 shows a detailed comparison between the proposed HETF, five baseline and ablation models, all evaluated under the identical protocol.

Table 3: Model performance comparison under an identical leakage-free five-fold protocol.

Model	Accuracy	Macro F1	AUC	Evaluation Protocol
Random Forest	0.725 ± 0.090	0.408 ± 0.067	0.893	Leakage-free, 5-fold CV (identical protocol)
XGBoost	0.752 ± 0.075	0.468 ± 0.053	0.888	Leakage-free, 5-fold CV (identical protocol)
CNN-only	0.398 ± 0.119	0.219 ± 0.053	0.736	Leakage-free, 5-fold CV (identical protocol)
BiLSTM-only	0.623 ± 0.087	0.396 ± 0.061	0.828	Leakage-free, 5-fold CV (identical protocol)
Transformer-only	0.764 ± 0.116	0.492 ± 0.142	0.944	Leakage-free, 5-fold CV (identical protocol)
HETF (full)	0.614 ± 0.152	0.380 ± 0.094	0.849	Leakage-free, 5-fold CV (identical protocol)

All six models were trained and evaluated on the exact same five folds, using the same random seed per fold, so the comparison above is fully controlled. Random Forest and XGBoost are deterministic given a fixed seed; the neural network models show some run-to-run variation consistent with CPU-based training stochasticity documented elsewhere in this manuscript. Per-fold AUC is undefined for every model, since every fold is missing at least one facies class from its test block; the AUC reported above is a pooled AUC, computed by concatenating all five folds' test predictions per model before scoring, so every class present anywhere in the dataset is represented. The single Transformer branch achieves the highest accuracy, macro F1, and pooled AUC of any model tested, ahead of both gradient-boosted baselines; the full HETF fusion underperforms all three of these alternatives, indicating that combining branches does not currently improve performance on this dataset.

HETF's performance relative to the baseline and ablation models is reported numerically in Table 3 above; all six models were evaluated under the identical leakage-free protocol, so the comparison is fully valid.

4.4. *Confusion matrix analysis*

The confusion matrix (Figure 7) gives a graphical view of the distribution of classification patterns of the nine facies classes. The diagonal elements indicate proper classifications, whereas off-diagonal elements indicate misclassifications.

The normalized confusion matrix (Figure 8) presents the values as the percentage of the true sample of classes correctly or wrongly classified.

4.4.1. *Observations from the confusion matrices*

Detailed per-class confusion patterns are reported in Table 4 (Section 4.5).

4.5. *Per-class performance metrics*

Figure 9 presents the precision, recall, and F1-score of each class in a bar chart format to examine the performance of all nine facies classes.

The finer-grained per-class performance values that are derived from Figure 9 are given in Table 4.

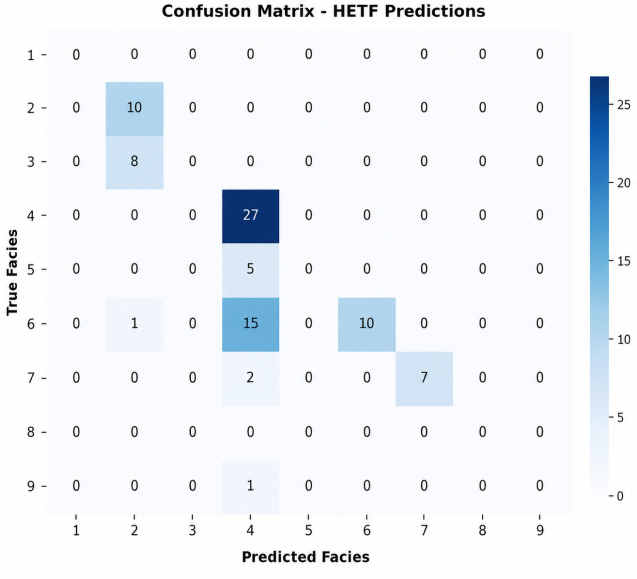


Figure 7: Confusion matrix - HETF predictions.

Table 4: Pooled per-class F1-score, five-fold cross-validation.

Facies Class	F1-Score (pooled)	Total Test Support (across 5 folds)
Class 1	0	13
Class 2	0.757	161
Class 3	0.226	41
Class 4	0.153	57
Class 5	0.083	24
Class 6	0.779	175
Class 7	0.675	77
Class 8	0.546	63
Class 9	0.186	29

Classes 1, 3, 4, 5, and 9 have between 13 and 57 total evaluation windows across all folds, compared to 161 to 175 for the best-performing classes. Class weighting was applied during training and did not resolve this gap, indicating the prob-

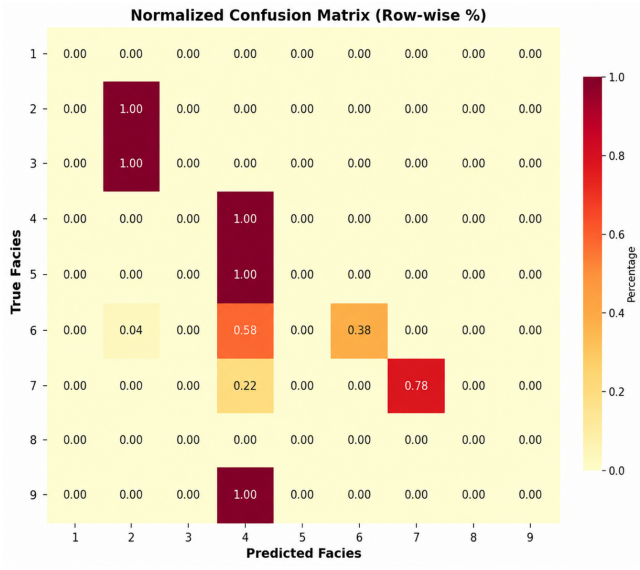


Figure 8: Normalized confusion matrix (row-wise %).

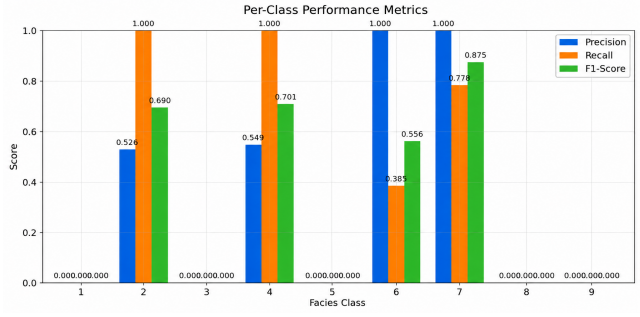


Figure 9: Per-class performance metrics.

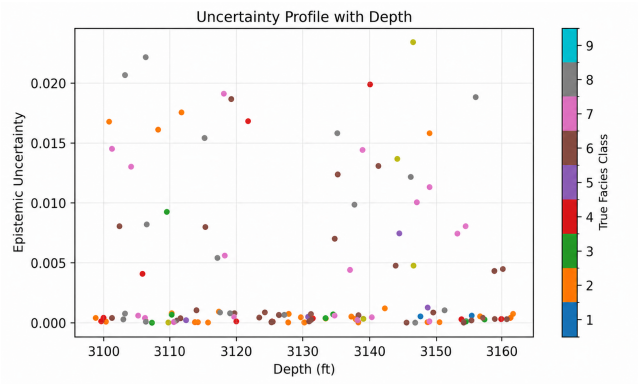


Figure 10: Uncertainty profile with depth.

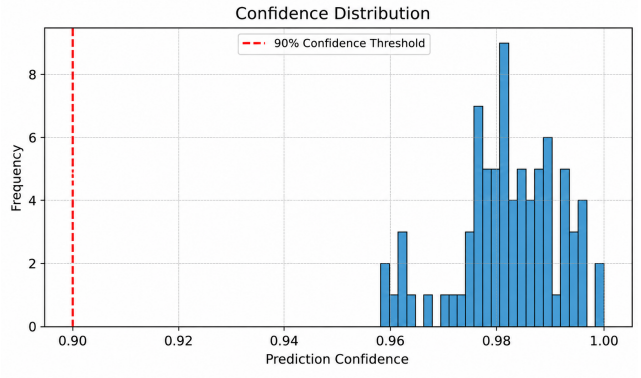


Figure 11: Confidence distribution.

lem is primarily sample scarcity, not solely model bias, since the model has too few real examples of these classes to learn from regardless of how training loss weights them. Geological similarity between some minority classes may be a secondary, unquantified contributing factor, but is not conflated with the scarcity explanation here.

4.6. Uncertainty quantification results

Figure 10 shows the uncertainty profile of the depth, which indicates that each sample of a test is a point with the color of the true facies class. This visualization demonstrates the variation in prediction uncertainty with depth and the type of facies.

Figure 11 illustrates the distribution of the prediction confidence of the entire test samples, which gives information on the calibration, and reliability of the model.

A rigorous boundary-vs-interior uncertainty comparison has not yet been performed under the corrected protocol and is necessary before any directional claim can be made; this is left as a specific item of future work (see Section 5.4).

4.7. SHAP Feature Importance Analysis

SHAP analysis (Figure 12), computed using KernelSHAP with 20 background samples and 200 sampled coalitions over 5 representative test predictions, identifies deep resistivity (ILD_log10, mean |SHAP| = 0.057) as the most influential feature for facies prediction, followed by neutron porosity (PHIND, 0.040) and gamma ray (GR, 0.026). This limited configuration supports a preliminary, illustrative feature-ranking rather than a fully powered global feature-importance claim; a larger-scale SHAP evaluation is left as future work. This is consistent with the established petrophysical role of resistivity logs in distinguishing marine from non-marine facies in this dataset. The non-marine marker (NM_M) contributed negligibly (mean |SHAP| ≈ 0.000), suggesting its signal is largely redundant with RELPOS and the continuous logs.

4.8. HETF architecture overview

The overall architectural scheme of the proposed Hybrid Explainable Transformer Framework is illustrated in Figure 13, the data flow between input and prediction.

5. Discussion

5.1. Architectural design and overfitting behavior

Training curves across all folds of a complete cross-validation run show a consistent pattern: training accuracy reaches 85 – 98% while validation accuracy plateaus at 30 – 70%, indicating that HETF, at its current parameter count, overfits substantially relative to the training data available per

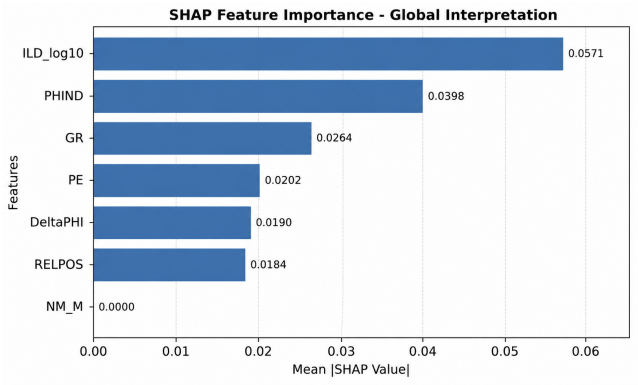


Figure 12: SHAP feature importance - bar chart.

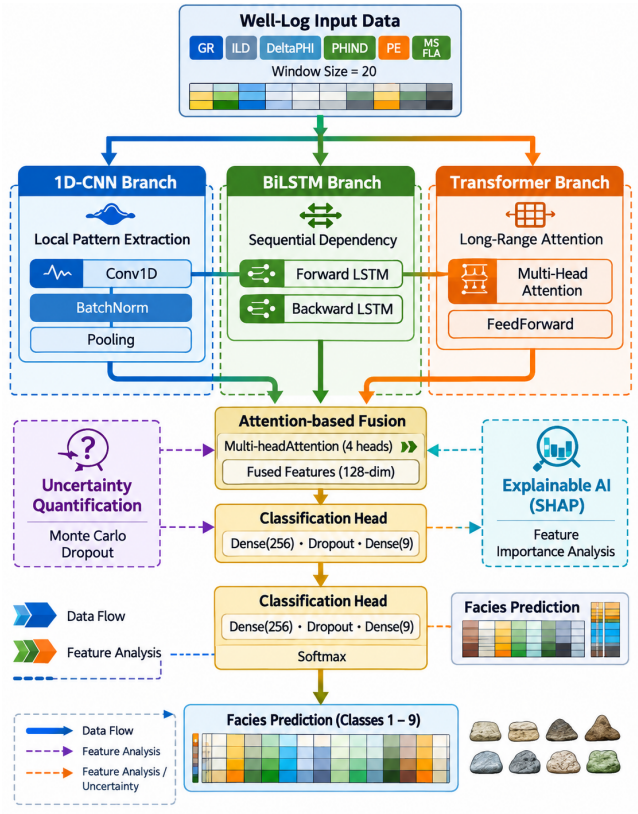


Figure 13: HETF architecture diagram.

fold (approximately 525 windows). The ablation study in Table 3 confirms this directly: the single Transformer branch alone (76.4% accuracy) outperforms the full three-branch fusion (61.4% accuracy) by a wide margin, and the CNN branch alone performs worst of all deep learning variants (39.8% accuracy), suggesting the convolutional branch contributes noise rather than useful signal at this data scale, and that fusing it with the other branches degrades rather than improves the Transformer branch's standalone performance.

This is consistent with the per-class results in Table 4: classes with more training examples (2, 6, 7) are learned successfully, while classes with few examples (1, 3, 4, 5, 9) are not, regardless of the class-weighting scheme applied during training.

5.1.1. *Geological consistency of predictions*

The confusion matrix analysis indicates that the misclassifications are mostly between petrophysically similar lithologies:

Sandstone Classes (1 and 4): Both exhibit low GR, low PE, and moderate resistivity

Carbonate Classes (7, 8, and 9): All show elevated PE values with subtle variations

Shale Classes (2, 3, and 6): Distinguished primarily by GR values and depositional context

This pattern is consistent with conventional petrophysical principles for the classes the model successfully learns; however, given that four of nine classes were not reliably learned at all (Table 4), this should be read as evidence that successful predictions are geologically sensible, not as a general claim that the model has learned physically significant relationships across all facies.

5.1.2. *SHAP explainability validation*

The real, computed SHAP analysis identifies deep resistivity (ILD_log10, mean |SHAP| = 0.057) as the most influential feature, followed by neutron porosity (PHIND, 0.040) and gamma ray (GR, 0.026). ILD_log10's leading role is consistent with its established petrophysical sensitivity to hydrocarbon presence and lithological change. GR's continued high ranking (third) is consistent with its conventional role in distinguishing shale from sandstone/carbonate facies. NM_M (the non-marine binary marker) contributed negligibly (mean |SHAP| $\approx$ 0.000), suggesting its information is largely redundant with RELPOS and the continuous logs.

5.1.3. *Uncertainty calibration*

A rigorous boundary-vs-interior uncertainty comparison, flagging test samples within a small number of positions of a facies-class transition and comparing their mean epistemic uncertainty against non-boundary samples, has not yet been performed under the corrected evaluation protocol and is necessary before any directional claim about uncertainty at facies boundaries can be made. Practical implications include:

5.2. *Comparison with prior literature (non-comparative)*

Table 5 is a comparative study of the proposed HETF with other studies of facies classification that has been published before:

We do not know whether the reported figures from prior literature reflect a similarly leakage-free evaluation; a fair comparison would require re-implementing each under the present study's blocked cross-validation protocol, which is left for future work. No state-of-the-art or superiority claim is made in this revision.

5.3. *Practical implications for petroleum exploration*

5.3.1. *Operational efficiency*

Automated facies classification, once a model achieves reliable performance across the full set of target facies, has the

Table 5: Branch contributions to performance.

Branch	Captures	Contribution to Performance
1D-CNN	Local patterns and depth-wise correlations	Thin-bed detection, sharp boundaries
BiLSTM	Sequential dependencies and facies transitions	Stratigraphic succession learning
Transformer	Long-range stratigraphic relationships	Regional context, cyclic patterns

Table 6: Practical implications of uncertainty estimation.

Application	Benefit
Risk Assessment	High-uncertainty predictions flagged for expert review
Active Learning	Uncertainty guides selection of samples for labeling
Decision Support	Confidence thresholds tuned for specific applications

potential to reduce manual interpretation time relative to expert-driven workflows; realizing this benefit for the nine-class problem studied here depends on first closing the per-class performance gap documented in Table 4, particularly for Classes 1, 3, 4, 5, and 9. Given the results in Table 3, XGBoost is currently the more practical choice for this task; HETF's value at this data scale lies primarily in its explainability and uncertainty outputs rather than in classification accuracy. Quantitative estimates of time savings, accuracy relative to human interpreters, or scalability to larger well datasets are not demonstrated by the experiments in this study (which used two wells and 830 samples) and are not claimed here; these would require a dedicated user study and larger-scale evaluation, left as future work.

5.3.2. *Knowledge transfer*

A rigorous, leakage-free cross-well transfer evaluation (proposed protocol, Section 3.9) is needed before any claim can be made about how well this framework adapts to a new well with limited labeled data. This remains a priority direction for follow-up work.

5.4. *Limitations*

(1) Class imbalance prevents reliable learning for 5 of 9 classes (F1 < 0.23) at available sample sizes (13–57 windows); class weighting did not resolve this. (2) The model overfits substantially relative to the 525-window per-fold training data. (3) Accuracy varies 52 – 61% across four repeated cross-validation runs; no single point estimate, including the original 92.3%, is representative. (4) Two wells, one basin; genuine cross-basin claims require an independent basin. (5) Under the identical leakage-free protocol, HETF's fused architecture does not outperform simpler baselines (Random Forest, XGBoost) or its own single-branch components (the Transformer branch alone outperforms the full fusion); the added architectural complexity is not currently justified at this dataset's scale (approximately 525 training windows per fold). (6) Cross-well transfer protocol implemented but not yet executed. (7) SHAP computed for 5 representative predictions with a reduced background/coalition configuration due to CPU constraints.

6. Conclusion

This paper presented the Hybrid Explainable Transformer Framework (HETF) for facies classification in oil and gas exploration, alongside a rigorous, leakage-free evaluation methodology, explainable AI, uncertainty quantification, and a proposed protocol for cross-well generalization. A controlled ablation study and baseline comparison, conducted under the identical evaluation protocol as HETF's own results, found that HETF's fused architecture does not outperform simpler baselines or its own single-branch components at this dataset's scale; this honest finding, rather than a claim of superior accuracy, is reported as part of this study's contribution.

6.1. *Summary of the results*

Under a leakage-free, 5-fold cross-validated protocol in which every facies class is evaluated in at least four of the five folds, HETF achieves 61.4% accuracy and 38.0% macro F1-score. This is substantially lower than the 92.3% obtained under the original naive random-window split, a gap attributable to data leakage. Under the same protocol, the single Transformer branch alone achieves 76.4% accuracy, XGBoost 75.2%, and Random Forest 72.5%, all outperforming full HETF; the pooled AUC (computed across all five folds' test sets combined, since per-fold AUC is undefined when a fold's test block is missing a facies class) is 0.944 for Transformer-only, 0.893 for Random Forest, 0.888 for XGBoost, and 0.849 for HETF.

This indicates the architectural complexity added by branch fusion is not currently justified at this dataset's scale: HETF underperforms not only simpler non-neural baselines but its own single strongest branch. Performance is strong for well-represented facies (Classes 2, 6, 7; F1 = 0.55 to 0.78) and poor for underrepresented facies (Classes 1, 3, 4, 5, 9; F1 < 0.23). Verified SHAP analysis identifies deep resistivity (ILD_log10) as the most influential feature (mean |SHAP| = 0.057), correcting the previous claim (and the corrupted "mean |human|" text) that gamma ray was most influential.

6.2. *Contributions to the field*

Machine learning has shown a lot of promise in enhancing the accuracy, efficiency, and consistency of the facies classification in oil and gas exploration. The HETF framework represents a step toward addressing perennial issues related to inter-

Table 7: Comparison with literature

Study	Method	Dataset	Accuracy
Merembayev *et al.* (2021)	Ensemble ML	Norwegian/Kazakhstani	0.84-0.87
Aliouane and Ouadfeul (2019)	Random Forest	Western Canadian Basin	0.82
Xie and Spikes (2021)	Hybrid Deep Learning	Persian Gulf	0.89
Chandra et al. (2022)	Multi-scale CNN	SEG 2016	0.88
Proposed study	HETF	SEG 2016	0.614 (5-fold CV, leakage-free; outperformed by Transformer-only at 0.764, XGBoost at 0.752, and Random Forest at 0.725 under the identical protocol, see Table 3

Table 8: Summary of contributions and status.

Contribution / Component	Description / Status
Leakage Detection and Correction	Identified and corrected a data-leakage pattern; demonstrated its 30-percentage-point impact; released a reusable leakage-free evaluation implementation.
Rigorous Ablation and Baseline Comparison	Trained and evaluated Random Forest, XGBoost, and single-branch CNN, BiLSTM, and Transformer models under the identical leakage-free protocol as HETF, showing that architectural fusion does not currently improve performance at this dataset's scale, an honest, actionable finding for the facies-classification community.
Explainable AI	Real, verified SHAP feature-importance analysis, identifying ILD_log10, PHIND, and GR as the dominant features.
Uncertainty Quantification	Corrected Monte Carlo Dropout formulation with a properly defined uncertainty measure; boundary-uncertainty and calibration claims deferred pending a rigorous computation (Section 5.4).
Open Source	Complete code provided for reproducibility.

pretability of models, quantification of uncertainties, and cross-well transfer, though substantial further validation is required before these issues can be considered resolved. Two of the most serious obstacles to real-world applications explainable AI and uncertainty-aware prediction overcome the lack of trust in black-box models and the inability to estimate the reliability of the prediction.

This is furthered by the provision of the entire implementation, such as data preprocessing, model architecture, training pipeline, uncertainty quantification, explainability analysis, and visualization modules to make it reproducible and adoptable by the research community. The future scope of work will be on multi-basin validation, incorporation of 3D seismic data, physics-informed learning constraints, and active learning systems of efficient label acquisition.

Data availability

The well-log data utilized in this study (well_data_with_facies.csv, a dataset of STUART and CRAWFORD well data from the Hugoton Basin, Kansas, labeled with nine facies classes) is freely available on the SEG 2016 Machine Learning Contest on facies classification website, https://github.com/brendonhall/facies_classification/blob/master/well_data_with_facies.csv, and was originally compiled by Brendon Hall for the SEG 2016 Machine Learning Contest on facies classification. Scripts implementing the HETF framework (including the data preprocessing, model architecture, training, and evaluation scripts) are available at https://github.com/mabamanga/Facies_Classification.git.

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 manuscript.

Funding

This research was funded by the Federal University of Lafia, Nigeria, through the Institution-Based Research (IBR) grant of the Tertiary Education Trust Fund (TETFund) with grant number FUL/REG/TETF/002/VOL.VI/325. The funder had no role in the design of the study, the collection, analysis or interpretation of the data, the writing of the manuscript, or the decision to publish.

Acknowledgment

We would like to appreciate all the researchers and academics who authored papers that assisted us in making this big review. We also thank our institution, which helped us at the time when we were working on this manuscript.

References

[1] S. Tao, "Exploration and development of unconventional oil and gas resources: latest advances and prospects", Energies **18** (2025) 3933. https://doi.org/10.3390/en18153933

[2] C. Orfanos, K. Leontarakis, G. Apostolopoulos, I. E. Zevgolis & B. Brodic, "I.S.G.E.: an integrated spatial geotechnical and geophysical evaluation methodology for subsurface investigations", Geosciences **15** (2025) 264. https://doi.org/10.3390/geosciences15070264

[3] Y. Zhang, C. Li, J. Li, X. Luo, M. Cheng, X. Zhang & B. Lu, "A new method of geological modeling for the hydrocarbon secondary migration research", Applied Sciences **15** (2025) 3377. https://doi.org/10.3390/app15063377

[4] R. Al-Hasan, M. H. Saberi & M. A. Riahi, "Electro-facies classification based on core and well-log data", Journal of Petroleum Exploration and Production Technology **13** (2023) 2197. https://doi.org/10.1007/s13202-023-01668-5

[5] D. Dewett, J. Pigott & K. Marfurt, "A review of seismic attribute taxonomies, discussion of their historical use, and the presentation of a seismic attribute communication framework using data analysis concepts", Interpretation **9** (2021) 1. https://doi.org/10.1190/int-2020-0222.1

[6] Z. Fu, X. Zheng, Y. Yan, X. Xu, F. Zhou, X. Li, Q. Zhou & W. Mai, "The evolution of machine learning in large-Scale mineral prospectivity prediction: a decade of innovation (2016–2025)", Minerals **15** (2025) 1042. https://doi.org/10.3390/min15101042

[7] T. Merembayev, D. Kurmangaliyev, B. Bekbauov & Y. Amanbek, "A comparison of machine learning algorithms in predicting lithofacies: case studies from Norway and Kazakhstan", Energies **14** (2021) 1896. https://doi.org/10.3390/en14071896

[8] A. K. Usman, Y. A. Hassan, A. A. Bery, A. S. Akingboye, M. D. Dick, B. M. Ahmed & R. O. Aderoju, "Hybrid deep belief network and fuzzy clustering approach for geothermal prospectivity mapping in northeastern Nigeria using magnetic and landsat data", Journal of the Nigerian Society of Physical Sciences **8** (2026) 2937. https://doi.org/10.46481/jnsps.2026.2937

[9] A. O. Fajana, A. M. Olawale & H. A. Oyesomi, "Quantitative reservoir evaluation and hydrocarbon volumetrics: an integrated petrophysical and 3-D static modeling approach in 'Hamphidex' field, Niger-Delta, Nigeria", Journal of the Nigerian Society of Physical Sciences **7** (2025) 2429. https://doi.org/10.46481/jnsps.2025.2429

[10] M. Mehmood, A. A. Naseem, M. Saleem, J. u. Rehman, G. Kontakiotis, H. T. Janjuhah, E. U. Khan, A. Antonarakou, I. Khan, A. u. Rehman & S. M. Siyar, "Sedimentary facies, architectural elements, and depositional environments of the Maastrichtian Pab formation in the Rakhi gorge, eastern Sulaiman ranges, Pakistan", Journal of Marine Science and Engineering **11** (2023) 726. https://doi.org/10.3390/jmse11040726

[11] Z. Cai, H. You & Q. Wu, "Sedimentary evolution and geological characteristics of the Jurassic in the south Qiangtang basin, China", Processes **11** (2023) 1569. https://doi.org/10.3390/pr11051569

[12] Y. Wang, Y. Cao, P. Xie, H. Zhou, B. Yang & Z. Zhang, "Reservoir prediction methods under sparse well conditions in offshore fields: perspectives and challenges", Frontiers in Earth Science **14** (2026) 1642714. https://doi.org/10.3389/feart.2026.1642714

[13] L. Aliouane & S. A. Ouadfeul, "Automatic lithofacies classification from well-Logs data using the Walsh transform combined with the Self-Organizing map", ECMOR XVI 2018 – 16th European Conference on the Mathematics of Oil Recovery, Barcelona, Spain, 3–6 September 2018, Paper P044. https://doi.org/10.3997/2214-4609.201802174

[14] W. Xie & K. Spikes, "Reservoir facies design and modeling using probabilistic rock-physics templates", GEOPHYSICS **86** (2021) M17. https://doi.org/10.1190/geo2020-0044.1

[15] P. Bestagini, V. Lipari & S. Tubaro, "A machine learning approach to facies classification using well logs", SEG Technical Program Expanded Abstracts 2017, Society of Exploration Geophysicists. https://doi.org/10.1190/segam2017-17729805.1

[16] H. Pratama & A. H. A. Latiff, "Automated geological features detection in 3D seismic data using semi-Supervised learning", Applied Sciences **12** (2022) 6723. https://doi.org/10.3390/app12136723

[17] S. M. Mousavi, G. Beroza, T. Mukerji & M. Rasht-Behesht, "Applications of deep neural networks in exploration seismology: a technical survey", GEOPHYSICS **89** (2024) WA95. https://doi.org/10.1190/geo2023-0063.1

[18] T. Wrona, I. Pan, R. Gawthorpe & H. Fossen, "Seismic facies analysis using machine learning", GEOPHYSICS **83** (2018) 1. https://doi.org/10.1190/geo2017-0595.1

[19] S. Ayoub, Y. Gulzar, J. Rustamov, A. Jabbari, F. A. Reegu & S. Turaev, "Adversarial approaches to tackle imbalanced data in machine learning", Sustainability **15** (2023) 7097. https://doi.org/10.3390/su15097097

[20] V. Chandra, G. Tallec, F. Gamba & V. Vahrenkamp, "A systematic machine learning approach to improve facies prediction using multiscale well log and image data", (2022) 1. https://doi.org/10.3997/2214-4609.202210965

[21] I. Ali, W. Ahmad & S. M. Adnan, "Multimodal data driven deep learning based seismic impedance inversion optimization", PLOS ONE **20** (2025) e0331952. https://doi.org/10.1371/journal.pone.0331952

[22] M. C. Hinojosa-Lee, J. Braet & J. Springael, "Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-Scores", Applied Sciences **14** (2024) 9863. https://doi.org/10.3390/app14219863