https://doi.org/10.31436/iiumej.v27i3.4355

Pipe-to-Soil Potential State Classification For Equatorial Gas Pipeline ICCP Systems Under Geomagnetically Induced Current Influence Using XGBoost Algorithm

MAIZURA MOHD SANI[1], AHMAD IHSAN MOHD YASSIN[1], MOHAMAD HUZAIMY JUSOH[1,3], NURHANI KASUAN[2], NUR DALILA KHIRUL ASHAR[2], MUHAMMAD ASRAF HAIRUDDIN[2]

[1] *Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Selangor Malaysia*
[2] *Faculty of Electrical Engineering, Universiti Teknologi MARA, Pasir Gudang Campus, Masai, Johor, Malaysia*
[3] *MAGDAS/CPMN Research Group, International Center for Space and Planetary Environmental Science, Kyushu University, Japan*

**Corresponding author: maizura23@uitm.edu.my*

(Received: 17 April 2026; Accepted: 15 July 2026; Published online: 12 September 2026)

ABSTRACT: Pipeline corrosion poses critical challenges for the oil and gas industry, resulting in economic losses and safety risks to humans, assets, and the environment. To mitigate corrosion, the Impressed Current Cathodic Protection (ICCP) system is employed, which typically maintains the pipe-to-soil potential (PSP) within the optimal protected range of -0.85 V to -1.2 V. However, geomagnetically induced currents (GICs) associated with geomagnetic storms can reduce the system's efficiency, causing the pipeline to be either underprotected or overprotected by the ICCP. In equatorial regions, this interference is quite challenging, as the Equatorial Electrojet (EEJ) can enhance daily daytime geomagnetic variability and may increase GIC-related interference in susceptible pipelines. This study proposes an Extreme Gradient Boosting (XGBoost) classification model to classify three gas pipeline ICCP protection states, i.e., protected, underprotected, or overprotected, at two equatorial pipeline locations, which is trained using ICCP and GIC data. The proposed model achieves robust classification performance, with accuracies of 94.22% at Location 1 and 86.50% at Location 2, and weighted F1-scores of 0.9432 and 0.8662, respectively. The model also demonstrates strong discrimination, with ROC-AUC values of 0.9922 and 0.9738 across sites.

ABSTRAK: Kakisan paip merupakan cabaran kritikal industri minyak dan gas. Ini mengakibatkan kerugian ekonomi serta risiko keselamatan kepada manusia, aset, dan alam sekitar. Bagi mengurangkan kakisan, sistem Perlindungan Katodik Arus Paksa (ICCP) digunakan, lazimnya mengekalkan potensi paip-ke-tanah (PSP) dalam julat perlindungan optimum antara -0.85V hingga -1.2V. Walau bagaimanapun, kecekapan sistem ini dipengaruhi oleh Arus Teraruh Geomagnet (GIC), yang berkait dengan ribut geomagnet dan boleh menyebabkan saluran paip mengalami perlindungan tidak mencukupi atau perlindungan berlebihan oleh ICCP. Di kawasan khatulistiwa, gangguan ini menjadi lebih mencabar kerana Elektrojet Khatulistiwa (EEJ) meningkatkan variasi geomagnet harian pada waktu siang dan berpotensi menambah gangguan berkaitan GIC pada saluran paip yang terdedah. Kajian ini mencadangkan model pengelasan berasaskan Extreme Gradient Boosting (XGBoost) bagi klasifikasi tiga status perlindungan ICCP saluran paip gas, iaitu dilindungi, kurang dilindungi, dan terlebih dilindungi, di dua lokasi saluran paip kawasan khatulistiwa menggunakan data ICCP dan GIC. Model yang dicadangkan mencapai prestasi pengelasan tinggi, dengan ketepatan sebanyak 94.22% di Lokasi 1 dan 86.50% di Lokasi 2, serta skor F1 berwajaran masing-masing 0.9432 dan 0.8662. Model ini juga menunjukkan keupayaan diskriminasi kukuh, dengan nilai ROC-AUC sebanyak 0.9922 dan 0.9738 bagi kedua-dua lokasi. Model ini boleh diguna sebagai sistem sokongan keputusan bagi pengendali saluran paip, membolehkan strategi penyelenggaraan proaktif sistem perlindungan katodik di kawasan khatulistiwa.

KEYWORDS: *Corrosion, Impressed current cathodic protection (ICCP), Geomagnetically induced current (GIC), Equatorial gas pipeline, XGBoost classification*

1. INTRODUCTION

Underground steel pipelines are the primary infrastructure for transporting oil and gas over long distances and serve as a main energy supply system for industrial, commercial, and residential sectors worldwide. One major challenge for these pipelines is corrosion, which can lead to pipe leakage, economic losses, safety risks, and environmental damage. To control corrosion, pipeline coatings and Impressed Current Cathodic Protection (ICCP) systems are commonly used together. Mimicking the electrochemical process, the ICCP system aims to ensure the pipeline acts as a cathode, reducing anodic oxidation reactions that cause corrosion [1]. As shown in Fig. 1, the ICCP system consists of anode beds of inert materials installed near the pipeline. Direct current (DC) supplied by a Transformer Rectifier Unit (TRU) is discharged through the anode bed into the surrounding soil, providing protective current to the pipeline and cathodically polarising the steel surface, thereby suppressing anodic metal dissolution and mitigating external corrosion. Pipeline protection is typically adequate when the pipe-to-soil potential (PSP), a monitoring ICCP parameter, is maintained between -850 mV and -1200 mV [2]. However, if the PSP is excessively negative, the ICCP overprotects the pipeline, which can lead to detrimental effects such as hydrogen embrittlement and damage to the pipeline coating [3].

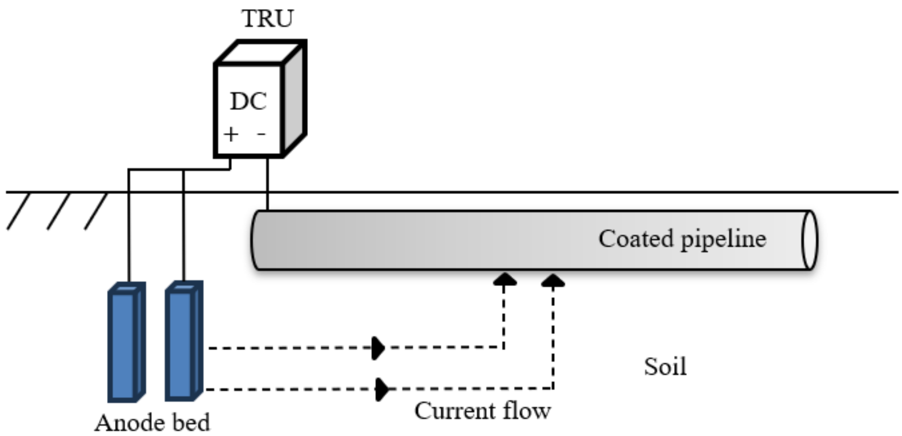

Figure 1. Schematic of the impressed current cathodic protection (ICCP) system for a buried pipeline

One factor that leads to overprotection and underprotection of the ICCP systems is the presence of stray telluric currents, or geomagnetically induced currents (GICs), which can be observed as fluctuations in PSP. Originating from space weather activity, the GIC hazard to the ICCP system has historically been associated with high-latitude regions due to the influence of the auroral electrojet near the pole [4]. However, the risk is now recognized as a global concern.. Pipelines in mid- and low-latitude regions have experienced out-of-safe-range PSP conditions during geomagnetic storm events [5] [6]. In Equatorial regions, pipelines experience more dynamic GIC interference. Equatorial pipelines experience interruptions from geomagnetic storms and interference from the Equatorial Electrojet (EEJ), a regular diurnal phenomenon caused by daily variations in Earth's magnetic field during solar-quiet (Sq) time. [7]. This indicates that pipelines in this region are subjected to slowly varying GICs every day. Although the magnitude of this daily interference may be less extreme than during a major storm, its constant and prolonged nature challenges the ICCP system's ability to operate efficiently. This is because the EEJ continuously fluctuates the PSP, affecting the ICCP system and potentially accelerating long-term corrosion at coating defects [7].

Machine learning has been increasingly applied to pipeline integrity monitoring and corrosion protection management, with early efforts focusing on physics-based and statistical models before transitioning to data-driven approaches. Physics-based modeling approaches, such as the Distributed-Source Transmission Line Theory (DSTL) model and equivalent circuit models [8], have been adopted to identify pipeline locations affected by GIC and have been applied in continuous monitoring systems, such as the nowcasting service in Finland [9]. While DSTL models simulate GIC propagation along pipeline networks, their computational complexity and reliance on geological conductivity profiles pose challenges for extensive pipeline systems. As an alternative, regression-based approaches using a Multilayer Perceptron (MLP) and a Nonlinear Autoregressive with Exogenous Inputs (NARX) model have been demonstrated as a reliable method for predicting PSP values in equatorial pipeline systems under GIC influence[10], showing that geomagnetic parameters such as dH/dt and the EEJ index carry significant predictive signal for PSP behavior. However, regression output provides continuous potential estimates rather than direct categorical protection status, which limits its utility as an actionable decision-support tool for pipeline operators. Rossouw and Doorsamy [11] presented one of the first attempts to classify PSP protection states, specifically the three operational categories of protected, underprotected, and overprotected for a pipeline near the South African region using historical ICCP system data. While their framework demonstrated the feasibility of data-driven PSP classification, the model's performance reportedly degraded significantly in the presence of stray current interference because the feature set was limited to ICCP system parameters and did not incorporate stray currents.

Despite this progress, a gap persists in the literature. No prior study has classified PSP protection states using a data-driven approach that explicitly incorporates GIC-related geomagnetic features as model inputs. This gap is particularly significant in the equatorial context, where the EEJ introduces a persistent, diurnal source of GIC-related PSP interference that operates independently of geomagnetic storm activity. Studies on GIC effects in mid- and low-latitude pipeline infrastructure have been reported [5][6]. However, these works documented telluric current interference without a machine learning framework for automated protection-state classification. Furthermore, no existing classification framework has been validated specifically for equatorial pipeline systems where the combined influence of EEJ-driven quiet-time variability and storm-time GIC disturbances creates a uniquely challenging operational environment for ICCP systems.

This study addresses the identified gap by proposing the first GIC-PSP classification framework for equatorial gas pipelines, implemented using the Extreme Gradient Boosting (XGBoost) algorithm [12], which is efficient at handling missing values and imbalanced data, both of which are characteristics of this study's ICCP pipeline data [13]. The framework incorporates dH/dt and the EEJ index (EUEL) as geomagnetic feature inputs alongside ICCP system data, addressing the limitation of ICCP-only approaches that resulted in reduced accuracy under stray-current conditions [11]. The following section details the methodology, including XGBoost model development and hyperparameter tuning, followed by the

https://doi.org/10.31436/iiumej.v27i3.4355

experimental results, performance evaluation, discussion, conclusions, and future research directions.

2. METHODOLOGY

To classify the PSP, this work proposes a four-stage methodology, as illustrated in Fig. 2. First is the data collection stage, which involves acquiring raw time-series sensor data from the pipeline's ICCP system, the PSP, and the TRU output current (Iout). Other data include the telluric current, or GIC, with indices of dH/dt, and the EEJ index (EUEL). The next stage is data preprocessing, which prepares the raw data by imputing missing values, resampling to ensure consistency, and extracting temporal features with specific time lags. The next stage is model training and optimization, where the data is split into training and test sets using a walk-forward validation scheme. The XGBoost model is then trained, and its hyperparameters are optimized using random search to find the best parameters. Finally, the performance metrics evaluate the model on the test data using accuracy, F1-score, the Receiver Operating Characteristic (ROC) curve, and the confusion matrix to confirm its reliability and effectiveness.

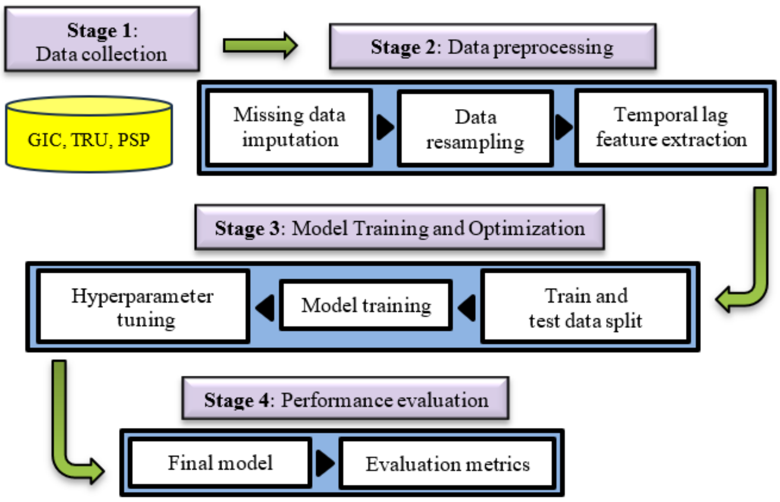

Figure 2. Block diagram of the ICCP classification system status using the XGBoost model

2.1. Data collection

The ICCP system PSP measurements were captured using the Close Interval Potential Survey (CIPS) [14] method. The pipeline's true potential, or off-potential, is measured over a 10-day period, from October 2, 2022, to October 12, 2022, using a data acquisition unit (DAQ) on two test posts at each location. These DAQs consist of voltage and current sensors integrated with a microcontroller, a real-time clock (RTC), and an SD memory card to record PSP and TRU output current. Both datasets have a sampling interval of 1 millisecond. Figure 3 shows the ICCP system and the CIPS measurement implemented in this study.

For this study, the PSP measurements were categorized into a three-class variable, 'underprotected', 'protected', and 'overprotected', based on the voltage thresholds as tabulated in Table 1. For the underprotected class, the dataset comprised 5,527 samples (27.42%) at Location 1 and 7,937 samples (39.38%) at Location 2. For the protected class, the corresponding numbers were 12,175 samples (60.39%) and 9,334 samples (46.30%), respectively. For overprotected samples, 2,458 (12.19%) were observed at Location 1 and 2,889 (14.32%) at Location 2. Figure 4 shows the plot of the normalized PSP values for this

https://doi.org/10.31436/iiumej.v27i3.4355

study. This distribution shows the operational behavior of the ICCP system, where the protected status corresponds to the dominant operational state during normal cathodic protection performance [11].

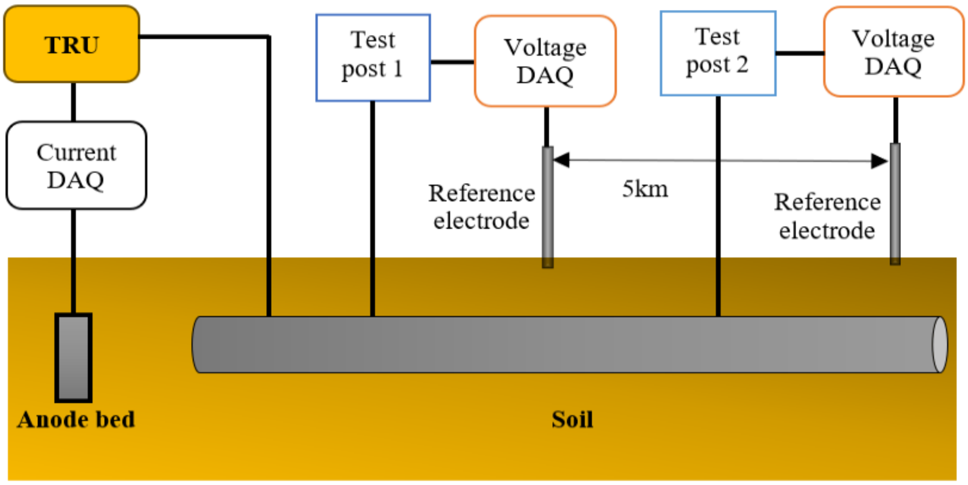

Figure 3. Data collection for PSP, TRU Iout at two pipeline locations at test posts 1 and 2

Table 1. PSP threshold of protected, underprotected and overprotected ICCP status

Characteristic	PSP value (V)
Protected	-0.85 to -1.2
Underprotected	Greater than -0.85
Overprotected	Less than -1.2

GICs originate from space weather activity, such as geomagnetic storms from solar disturbances, that perturb the Earth's magnetosphere and cause variations in the Earth's magnetic field [15]. Based on Faraday's law, these rapid temporal variations in the Earth's magnetic field give rise to geoelectric fields at the Earth's surface, which then drive GICs. In this study, the time derivative of the horizontal Earth's magnetic field, i.e., dH/dt represents a proxy for GIC activity, specifically in the low-latitude region [16]. According to established space weather studies, dH/dt is strongly correlated with the intensity of induced electric fields and has been adopted as a severity level for GICs in ground-based technological systems [17]. To compute dH/dt, the H-component was obtained from the magnetic observatory station at Universiti Teknologi MARA (UiTM), Pasir Gudang, and then numerically differentiated. In equatorial regions, the EEJ is a concentrated eastward electric current flowing in the ionospheric E-layer along the magnetic dip equator. We used the EEJ index (EUEL) from Kyushu University, Japan, to quantify EEJ strength [18]. Fig. 4 shows plots of PSP at Locations 1 and 2, TRU Iout, dH/dt, and EUEL. All variables are min-max normalized to the range [0, 1] to synchronize the data. Fluctuations in PSP at both locations indicate GIC influence on pipeline protection states.

2.2. Data Preprocessing

2.2.1. Missing data imputation

The PSP and TRU datasets contained occasional short-term gaps due to sensor interruptions, with missing values ranging between 7.5% and 18.6%. We imputed these gaps using linear interpolation, which assumes a straight-line relationship between adjacent observations and preserves signal continuity. Given two consecutive observations, x_i and x_{i+1} a missing value, $\hat{x}(t)$ at time t, was estimated using Eq. (1):

$$\hat{x}(t) = x_i(t) + \frac{(t-t_i)}{(t_{i+1)}-t_i)}(x_{i+1} - x_i) \quad (1)$$

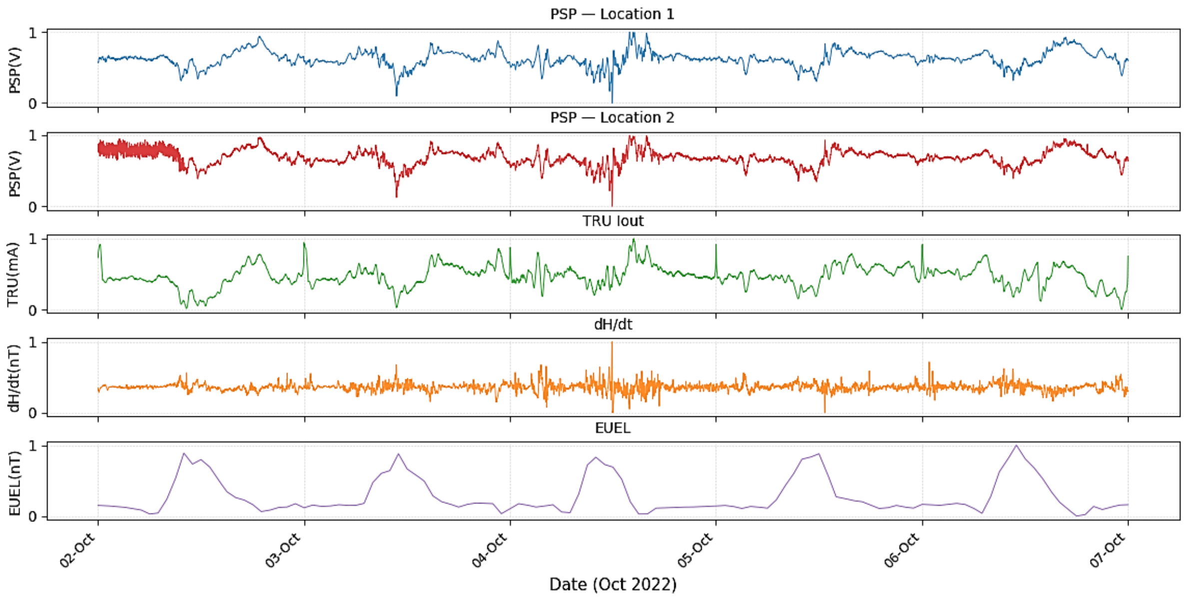

Figure 4. Normalized PSP for Locations 1 and 2, TRU Iout, dH/dt, and EUEL from October 2, 2022, to October 7, 2022

2.2.2. Data resampling

The PSP, TRU output current, and dH/dt signals were originally sampled at millisecond resolution. To harmonize the dataset sampling rate, all variables were resampled to a one-minute resolution (1/60 Hz) to synchronize with the EUEL data. Down-sampling was performed using a two-step procedure. First, a digital low-pass filter was applied to attenuate frequency components above the new Nyquist limit, followed by decimation by a factor of 60. The filtering stage employed an anti-aliasing FIR filter (MATLAB resample with the default Kaiser window design) to keep the signals band-limited before decimation.

2.2.3. Temporal lag feature extraction

Temporal lag features were extracted to capture variations and historical patterns across all input data, which improve prediction performance. For EUEL, dH/dt, and TRU, lagged features were constructed at 1, 5, 10, and 30 minutes. The lag intervals were selected to capture data at different timescales for short-term transients (1-5 minutes), medium-term trends (10 minutes), and extended variations (30 minutes). The temporal lag set can be expressed as in Eq. (2):

$$X = [v(t-\tau)] \text{ for } v \in \{EUEL, \frac{dH}{dt}, TRU\} \text{ and } \tau \in \{0, 1, 5, 10, 30\} \quad (2)$$

where *t* represents the current time step of 1-minute resolution.

2.3. Model training and optimization

2.3.1. Train and test data split

The dataset was split into training and test sets, labeled as blue and red boxes, respectively, as shown in Figure 5. We split the data using a day-grouped walk-forward validation scheme

[19] to reduce overfitting in time-series data. In this figure, chronological order prevents information leakage from future samples into the training set, providing an independent validation set for performance assessment. Five folds were constructed so that the training window expanded progressively over time, and each test fold consisted of the subsequent day's observations.

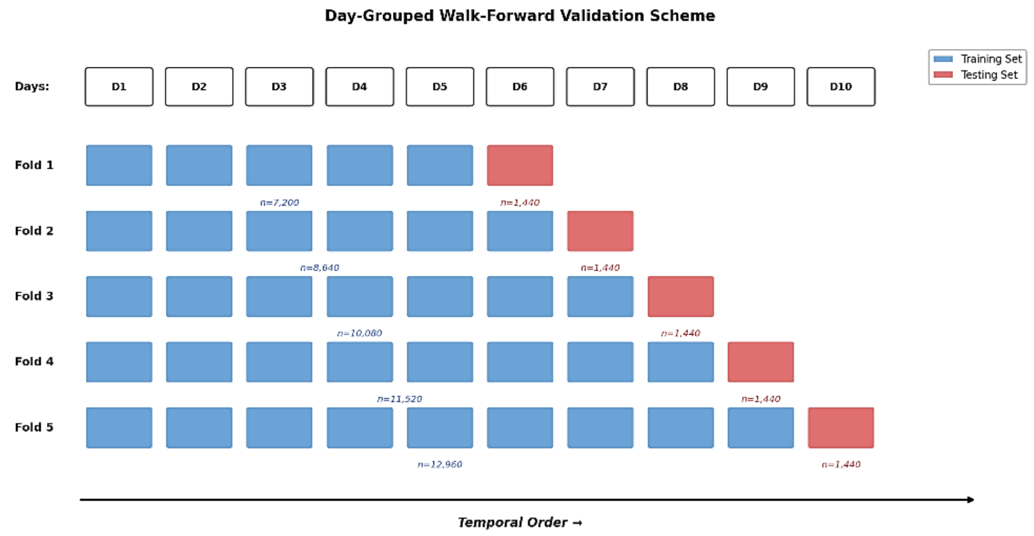

Figure 5. Day-grouped walk-forward validation

2.3.2. Model training

The XGBoost classifier was trained on the training partition using the selected hyperparameters. XGBoost learns an additive ensemble of decision trees by minimizing a regularised objective function, $\mathcal{L}(\phi)$ which balances data fit and model complexity. The overall objective function is defined as in Eq. (3):

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{t} \Omega(f_k) \tag{3}$$

where $l(\hat{y}_i, y_i)$ denotes the loss between the predicted output and the true label, and $\Omega(f_k)$ is the regularisation term that penalizes the complexity of the tree f_k. At iteration t, XGBoost adds a new tree, f_t by minimising $\mathcal{L}^{(t)}$ as in Eq. (4):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{k=1}^{t} \Omega(f_t) \tag{4}$$

The loss term, given by $\sum_{i=1}^{n} \boldsymbol{l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))}$, measures the discrepancy between the true label, $\boldsymbol{y_i}$ and the updated prediction obtained by adding the output of the new tree $\boldsymbol{f_t(x_i)}$, to the prediction from the previous $\boldsymbol{t-1}$ trees, $\boldsymbol{\hat{y}_i^{(t-1)}}$. The regularization term, $\boldsymbol{\Omega(f_t)}$ penalizes the complexity of the newly added tree at the iteration $\boldsymbol{t}$ to control model complexity and reduce overfitting.

Multiclass log-loss was selected as the evaluation metric because it penalizes overconfident incorrect predictions more severely than uncertain incorrect predictions. The histogram-based tree construction algorithm was implemented to reduce computational complexity by discretizing continuous features, enabling efficient training in large feature spaces while maintaining model accuracy.

2.3.3. Hyperparameter tuning

Hyperparameter tuning was implemented using a random search over 100 iterations, with each candidate configuration evaluated using five-fold time-series cross-validation and scored on weighted F1. The hyperparameter search space comprised the following values; maximum depth of 4, 6, 8, 10 and 12, minimum child weight with the value of 1, 3, 5, and 7, subsample ratio; 0.7, 0.8, 0.9, and 1.0, column sampling ratio; 0.7, 0.8, 0.9, and 1.0, gamma; 0.0, 0.1, 0.2, 0.3, L1 regularisation; 0.0, 0.001, 0.01, 0.1, L2 regularisation; 0.1, 0.5, 1.0, 5.0, and 10.0, learning rate with 0.01, 0.02, 0.03, and 0.05, and number of estimators of 1500 and 2000. Early stopping with a patience of 50 rounds was applied during each fold to prevent overfitting. The combination with the highest mean weighted F1 across folds was then used to train the final model.

2.4. Performance evaluation

Classification performance was evaluated using standard metrics, i.e., accuracy, macro and weighted F1-scores, and Receiver Operating Characteristic (ROC) analysis [20]. The weighted F1-score served as the main metric, weighted by each class's number of true instances, making it suitable for imbalanced multiclass data. ROC analysis using a one-vs-rest approach was conducted, and the Area Under the ROC Curve (AUC-ROC) was calculated for each class as a threshold-independent measure of performance.

3. RESULTS AND DISCUSSION

3.1. Model performance for 5-fold cross-validation

Figure 6 shows the results for accuracy, weighted F1-score, macro F1-score, and ROC-AUC from 5-fold cross-validation for both locations using a day-grouped walk-forward validation scheme. Location 1 outperformed Location 2 across all metrics, with higher mean values observed in each subplot. Performance at Location 1 was generally more consistent across folds, with smaller fluctuations. In contrast, Location 2 showed greater variability, particularly in fold 2, with lower values observed across all metrics. Despite this variation, ROC-AUC remained high across all folds at both locations, suggesting that the classifier maintained good class separability even when classification scores were lower. The fifth fold achieved the highest performance across all metrics at both locations, indicating that more training data progressively improves model generalization.

3.2. Performance per class

The one-vs-rest ROC curves are depicted in Fig. 7(left) for Location 1 and Fig. 7(right) for Location 2. Location 1 achieved high performance, with an AUC ranging from 0.986 to 0.997, whereas Location 2 achieved an AUC ranging from 0.940 to 0.995. The over-protected class consistently achieved the highest AUC at both locations (0.997 and 0.995 for Locations 1 and 2, respectively). Conversely, for both locations, the protected class demonstrated slightly lower discrimination performance. Despite these minor variations, the sustained high ROC-AUC performance across both locations validates the model's robustness for ICCP pipeline protection monitoring applications.

Fig.8 shows the confusion matrices for both monitoring sites. At Location 1, the model correctly identifies 94.57% (209/221) of under-protected samples, 95.25% (702/737) of protected samples, and 90.79% (138/152) of over-protected samples. At Location 2, the model correctly identifies 92.76% (346/373) of under-protected samples, 79.61% (449/564) of protected samples, and 92.65% (189/204) of over-protected samples. The protected class at Location 2 has the highest number of misclassifications, with 105 samples classified as under-protected and 10 as over-protected. This higher misclassification rate is attributed to the narrower margin between the protected (46.30%) and under-protected (39.38%) class distributions at Location 2 compared to Location 1 (60.39% vs 27.42%), which reduces inter-class separability during training.

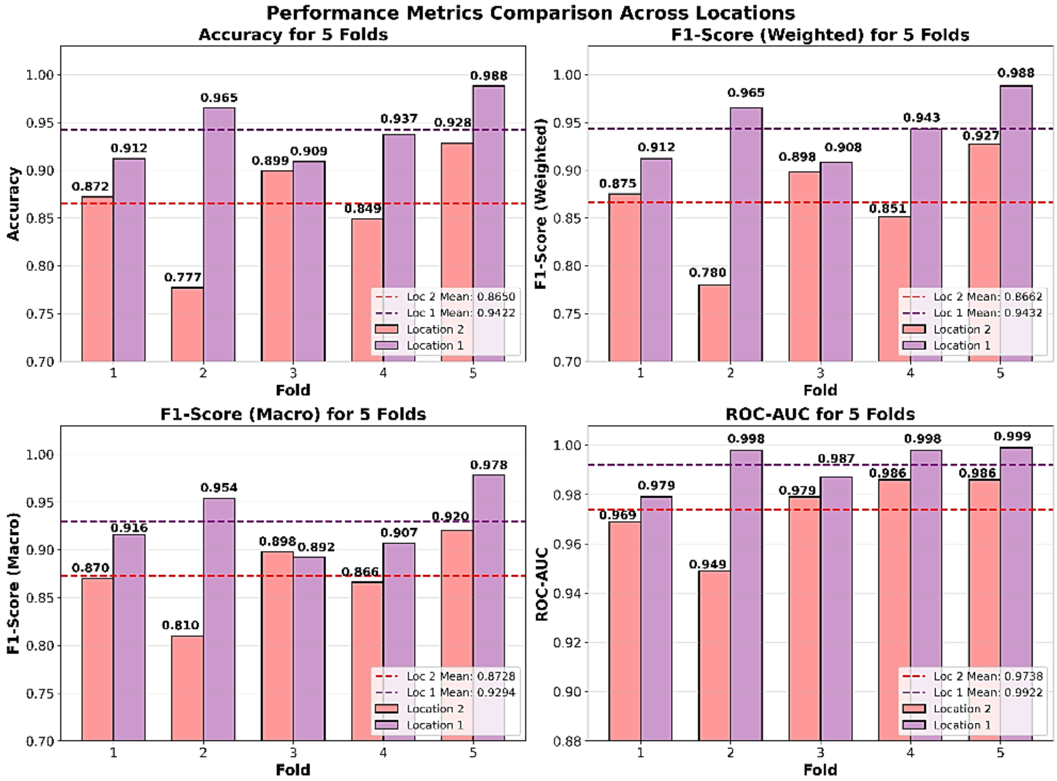

Figure 6. Model performance for five-fold cross-validation

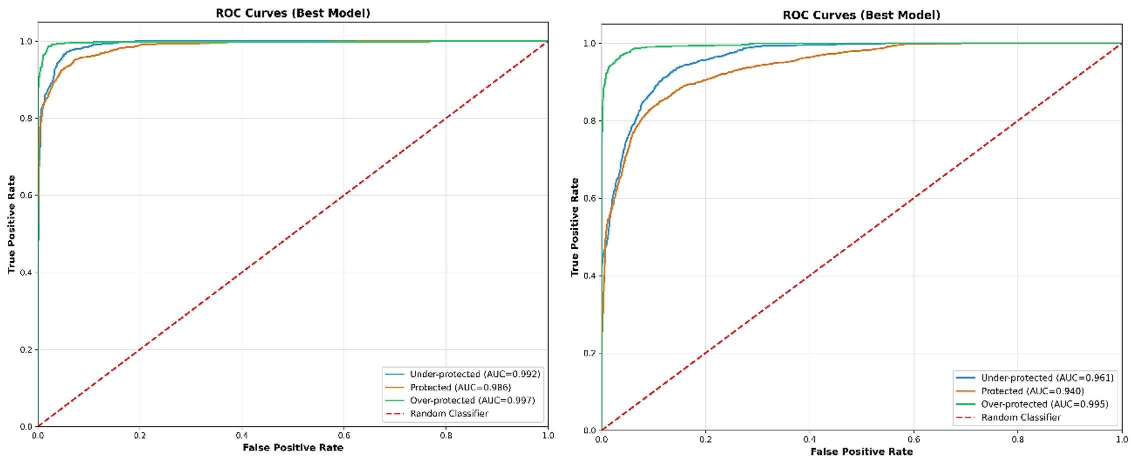

Figure 7. One-vs-rest ROC curves for three-class pipeline protection status classification for Location 1(left) and Location 2(right)

No under-protected sample is misclassified as over-protected, or vice versa, at both locations. This is attributed to the distinct separation of their PSP voltage ranges, with the entire protected zone (-0.85 V to -1.2 V) residing between them. As the protected samples lie in the intermediate range of the PSP spectrum, this class is more susceptible to boundary effects when readings fluctuate near either threshold, where PSP values near -0.85 V and -1.2 V introduce ambiguity in class assignment between adjacent classes.

https://doi.org/10.31436/iiumej.v27i3.4355

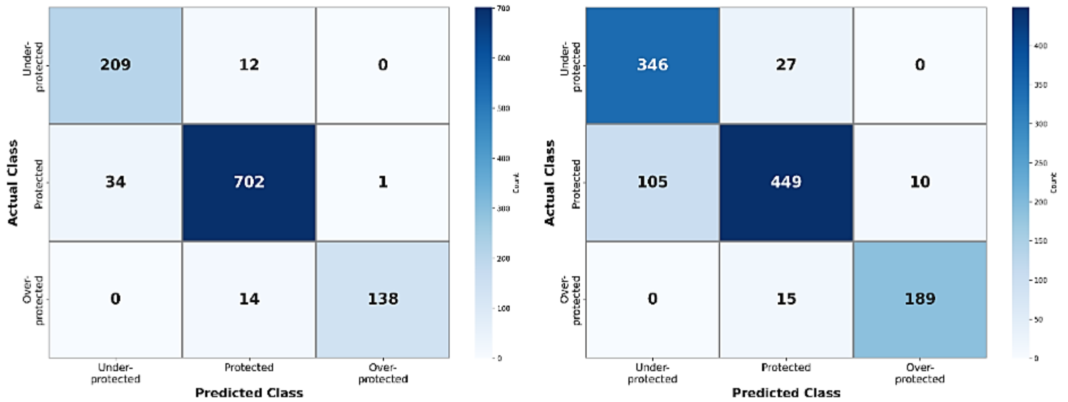

Figure 8. Confusion matrix of the ICCP system status classification (under-protected, protected, and over-protected) for Location 1(left) and Location 2(right)

3.3. Feature importance analysis

Fig. 9 presents the Shapley Additive explanations (SHAP) [21] global feature importance ranking for Locations 1 and 2. At both sites, the current (un-lagged, τ=0) values of EUEL and TRU dominate the ranking, occupying the top two positions: EUEL_lag0m ranks first at Location 1, while TRU_lag0m ranks first at Location 2, with the other swapping to second place. This indicates that the instantaneous EEJ intensity and TRU output current carry the strongest direct relationship with the current PSP protection state, consistent with their role as the most immediate drivers of ICCP system behavior.

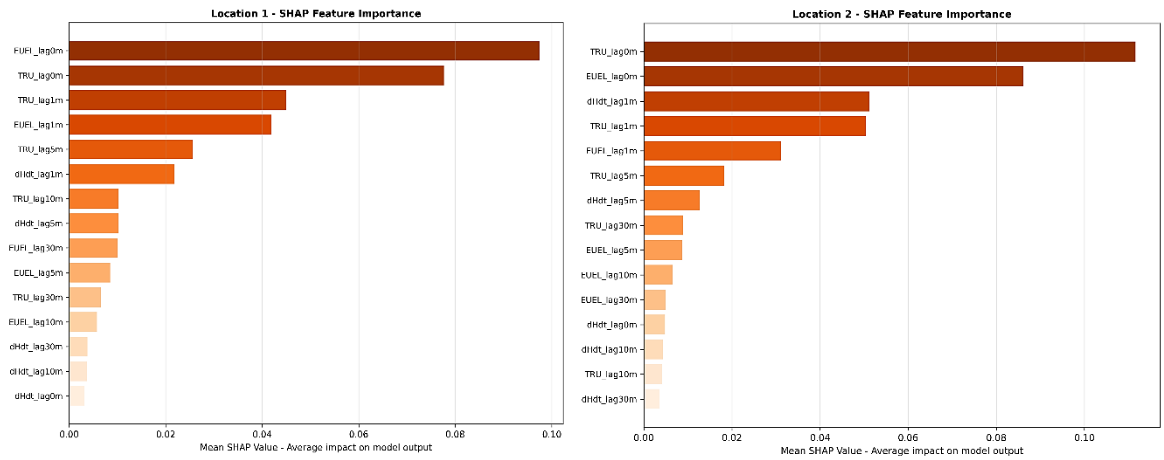

Figure 9. Global SHAP feature ranking for Location 1 and Location 2

A consistent pattern across both locations concerns dH/dt. Its current value (dHdt_lag0m) ranks lowest among all fifteen features at both sites, while its 1-minute lag (dHdt_lag1m) ranks substantially higher, sixth at Location 1 and third at Location 2. This contrast suggests that the instantaneous dH/dt reading carries comparatively little direct classification value, whereas the value recorded one minute prior is considerably more informative. The consistently lower SHAP importance of the instantaneous dH/dt value (dHdt_lag0m) relative to its one-minute lag (dHdt_lag1m) at both locations suggests a finite temporal delay between geomagnetic field variation and its manifestation in the PSP signal. This may be because the magnetic observatory at UiTM Pasir Gudang is not co-located with the pipeline monitoring sites, meaning the recorded dH/dt is regional rather than a direct local measurement of the geoelectric field at the pipeline [22]. This spatial offset may introduce a temporal lagged discrepancy between the observed dH/dt and the resulting PSP response. Beyond a 1-minute lag, the reduced importance of 10-minute and 30-minute lagged variables has been observed, which indicates that the system response is mainly affected by short-term dynamics within the first few minutes.

3.4. Contribution of GIC features on model performance

Table 2 shows XGBoost model performance using three input feature sets: TRU only, TRU with dH/dt, or TRU with dH/dt and EUEL. Using TRU current output alone as input features has exhibited limited classification performance at both monitoring locations. At Location 1, the accuracy, F1-macro, F1-weighted, and ROC-AUC were 0.6769, 0.6092, 0.6664, and 0.8572, respectively, while the corresponding values at Location 2 were 0.5360, 0.5451, 0.5207, and 0.7716. Although TRU current reflects the impressed current supplied to the pipeline, it does not capture transient geomagnetic disturbances that influence the PSP. This modest performance indicates that TRU current data alone is insufficient for predicting PSP protection states under GIC-disturbed conditions.

Among the evaluated feature combinations, including dH/dt yields the greatest performance improvement, with consistent gains across all metrics. The F1-macro increased by 0.2031 and 0.2146 at Locations 1 and 2, respectively, indicating improved class balance in prediction performance. The ROC-AUC values also increased to 0.9422 and 0.9194, demonstrating enhanced class separability. This improvement is physically justified, as dH/dt represents rapid geomagnetic fluctuations associated with GIC-induced disturbances that directly affect PSP behavior.

Table 2. Performance of the XGBoost model based on different input combinations.

Loc	Features	Accuracy	F1-scores (macro)	F1-scores (weighted)	ROC-AUC
1	TRU	0.6769	0.6092	0.6664	0.8572
	TRU, dH/dt	0.8346	0.8123	0.8381	0.9422
	TRU, dH/dt, EUEL	0.9422	0.9294	0.9432	0.9922
2	TRU	0.5360	0.5451	0.5207	0.7716
	TRU, dH/dt	0.7493	0.7597	0.7509	0.9194
	TRU, dH/dt, EUEL	0.8650	0.8728	0.8662	0.9738

Adding EUEL to form the complete feature set produced further improvements across all evaluation metrics, although the magnitude of improvement was smaller than that associated with dH/dt. At Location 1, the model achieved an accuracy of 0.9422, F1-macro of 0.9294, F1-weighted of 0.9432, and ROC-AUC of 0.9922. At Location 2, the corresponding values were 0.8650, 0.8728, 0.8662, and 0.9738. The consistent improvement for all metrics indicates that EUEL provides important information. As an indicator of EEJ strength, EUEL captures geomagnetic variations that dH/dt does not, therefore improving characterization of GIC-induced PSP perturbations.

3.5. Comparison of different classification model performance

Table 3 presents a comparative evaluation of the XGBoost classifier, which was performed using five commonly deployed machine learning algorithms, i.e., Random Forest (RF), LightGBM, CatBoost, Multilayer Perceptron (MLP) Artificial Neural Network (ANN), and Support Vector Machine (SVM), evaluated under identical day-grouped walk-forward cross-validation conditions. All comparison classifiers were hyperparameter-optimized using random search over a single chronological train-validation split (75/25), with 100 search iterations for RF, LightGBM, CatBoost, and MLP, and 15 iterations for SVM due to its higher computational cost. The XGBoost model achieved the highest classification accuracy at both

https://doi.org/10.31436/iiumej.v27i3.4355

monitoring locations, with mean accuracies of 94.22% and 86.50% at Locations 1 and 2, respectively, outperforming the next-best classifier (LightGBM) by approximately 6 percentage points at both sites. Among the comparison classifiers, LightGBM performed most competitively, achieving mean accuracies of 86.03% and 79.76% at Locations 1 and 2, respectively. RF and CatBoost performed moderately, while MLP-ANN showed greater variability. SVM performed worse than all methods at both locations, with mean accuracies of 49.22% and 45.42%, marginally above the theoretical random baseline for a three-class problem. XGBoost's consistent superiority across all metrics at both monitoring locations indicates that this algorithm is well suited to the GIC-influenced PSP classification task.

Table 3. Performance of the XGBoost model compared with other classification models.

Loc	Algorithm	Accuracy	F1-scores (macro)	F1-scores (weighted)	ROC-AUC
1	Random Forest	0.8390	0.8200	0.8405	0.9581
	Light GBM	0.8603	0.8443	0.8642	0.9613
	CatBoost	0.8306	0.8224	0.8377	0.9625
	MLP(ANN)	0.8147	0.7860	0.8182	0.9304
	SVM	0.4922	0.5374	0.4733	0.8699
	XGBoost	**0.9422**	**0.9294**	**0.9432**	**0.9922**
2	Random Forest	0.7312	0.8121	0.7312	0.9364
	Light GBM	0.7976	0.7546	0.7976	0.9509
	CatBoost	0.7143	0.7364	0.7143	0.9440
	MLP(ANN)	0.7835	0.7814	0.7835	0.9423
	SVM	0.4542	0.4390	0.4542	0.8512
	XGBoost	**0.8650**	**0.8728**	**0.8662**	**0.9738**

3.6. Computational Performance and Operational Considerations

Computational performance shows the speed at which this study's computer system can execute the XGBoost classification model, demonstrating efficiency suitable for near-real-time pipeline protection monitoring. All timing measurements were obtained using Python's high-resolution performance counter (time.perf_counter()). Timing measurements were recorded on a system equipped with an Intel Core i7-14700HX processor and 16 GB of RAM. Model training across five walk-forward cross-validation folds took approximately 11.12 seconds, with a mean training time of 2.22 seconds per fold. For the inference stage, in which the trained model classifies the instantaneous PSP protection state from incoming sensor features, it is computationally inexpensive, with a mean classification time of 0.032 milliseconds per sample, equivalent to approximately 31,000 classifications per second. This indicates that the model can process the input sensor data resolution within the available sampling window.

Table 4 compares mean training and inference times across all six classifiers evaluated in this study. All classifiers achieve inference times well below the sensor data's one-minute sampling resolution, indicating that real-time classification is computationally feasible for all methods evaluated. Among the tree-based ensemble methods, XGBoost and RF demonstrated comparable training times of 2.22 and 1.29 seconds per fold, respectively, while LightGBM was marginally slower at 2.41 seconds per fold. CatBoost exhibited the highest training time among tree-based methods at 19.19 seconds per fold. The MLP achieved the fastest inference time at 0.0014 ms per sample, followed closely by CatBoost at 0.0026 ms per sample. XGBoost's inference time of 0.032 ms per sample, while slightly higher than MLP and CatBoost, remains well within real-time operational requirements. SVM has the highest inference time at 0.221 ms per sample. XGBoost is therefore selected not for computational efficiency alone; rather, its high classification accuracy at both monitoring locations, combined

https://doi.org/10.31436/iiumej.v27i3.4355

with acceptable inference latency and low training time, justifies its adoption as the primary classifier.

Table 4. Mean training time and mean inference time for different classifier models

Classifier	Mean Training Time	Mean Inference Time
MLP (ANN)	1.909	0.0014
CatBoost	19.194	0.0026
LightGBM	2.406	0.0099
XGBoost	**2.224**	0.0321
Random Forest	1.285	0.0432
SVM	8.963	0.2207

With low computational overhead, this model's classification output could map directly to actionable decisions within the pipeline operator's maintenance workflow. Pipeline engineers monitoring the ICCP system can use the PSP output state as a continuous decision-support signal alongside conventional PSP readings. A protected classification indicates that the system is operating within the acceptable potential range of -0.85 V to -1.2 V and requires no immediate intervention. An underprotected classification alerts the engineer that the pipeline is at risk of accelerated corrosion and prompts actions such as increasing the TRU output current, inspecting for coating defects, or verifying the integrity of ICCP system components. An overprotected classification indicates excessive cathodic current, which may lead to hydrogen embrittlement or coating disbondment, and requires reducing TRU output current. In persistent or severe cases, particularly where recurrent underprotection is detected despite TRU adjustment, the classification output may serve as an early indicator warranting physical excavation and direct pipeline inspection. Integrating this classification framework into a complete real-time deployment architecture, including communication protocols and SCADA system interfacing, represents a direction for future work.

3.7. Limitations and future work

The present study spans approximately ten days under geomagnetically quiet conditions and a moderate storm event from 3 to 5 Oct 2022 (minimum EDst index value of -59 nT) [23]. As a result, the training data largely captures EEJ-driven diurnal variability but does not encompass strong or severe GIC behavior or seasonal variations in EEJ intensity. Because the observation window contains no major geomagnetic storm events, the model's performance under stronger geomagnetic forcing has not been evaluated. These constraints reflect a broader challenge in equatorial GIC research, where observational data from equatorial and low-latitude regions remain limited [24]. A similar challenge has been reported in machine-learning-based space weather prediction, where data scarcity and class imbalance in rare-event datasets have led researchers to employ strategies such as transfer learning, data augmentation, and resampling methods to improve model robustness and generalization [25]. In pipeline corrosion monitoring, collecting continuous, high-resolution field data is also constrained by cost, sensor access, and operational considerations, as interrupting the ICCP system to obtain PSP data is not feasible. Future work should prioritize acquiring longer-term, multi-season datasets covering varying levels of geomagnetic activity, including storm-time events, at different geographical sites to improve model robustness and validate generalization across a broader range of ICCP operational conditions. Additionally, data augmentation strategies such as conditional generative adversarial networks (cGAN) or synthetic minority oversampling may help address the inter-class separability and temporal coverage challenges identified in this study.

4. CONCLUSION

This study developed an XGBoost classification model incorporating temporal lag features derived from GIC-influenced geomagnetic field variations to address the limited availability of data-driven frameworks for classifying PSP protection states in an ICCP system for an equatorial-region gas pipeline under GIC disturbances. The proposed model achieved classification accuracies of 94.22% and 86.50% at Locations 1 and 2, respectively, with ROC-AUC values of 0.9922 and 0.9738, indicating strong discriminative capability across all three protection states at both sites. A comparative evaluation of five classifiers, i.e., Random Forest, LightGBM, CatBoost, MLP, and SVM, showed consistent superiority of XGBoost at both monitoring locations. XGBoost's consistent superiority across all metrics and both locations, combined with an inference time of 0.032 milliseconds per sample, indicates the feasibility of the proposed framework for real-time operational deployment within the sensor time-sampling window. Among the input features, dH/dt contributed the largest performance increment as the most informative variable for capturing GIC-induced PSP disturbances, while EUEL provided additional discriminative information that further refined the classification boundaries. The present study has certain limitations. The ten-day observation window predominantly captures EEJ-driven diurnal variability without encompassing strong or severe storm-time GIC events (which depend on space weather geomagnetic storm events) or seasonal variations in EEJ intensity. Furthermore, the dH/dt and EUEL features should be specific to locally acquired geomagnetic data near the pipeline. Future work should prioritize acquiring longer-term, multi-season datasets spanning varied geomagnetic activity levels, exploring data augmentation strategies to address inter-class separability challenges, and cross-regional validation of the proposed methodology for equatorial pipeline systems. The proposed approach highlights its potential as a reliable support tool for pipeline engineers, particularly for ICCP protection state monitoring in equatorial pipeline systems under the influence of geomagnetic storms and the EEJ phenomenon.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support provided by Universiti Teknologi MARA and the Ministry of Higher Education Malaysia.

REFERENCES

[1] Peabody AW. (2018) Peabody's Control of Pipeline Corrosion (3rd ed.). Houston, NACE International. https://doi.org/10.5006/37617

[2] Gummow RA. (2018) Cathodic protection criteria-A critical review of NACE standard RP-01-69. Mater. Perform., 57(1):28-35. https://doi.org/10.5006/MP2018_57_1-28

[3] Hosokawa Y, Kajiyama F, Nakamura Y. (2004) New cathodic protection criteria based on direct and alternating current densities measured using coupons and their application to modern steel pipelines. Corrosion, 60(3):304-312. doi:10.5006/1.3287735

[4] Gummow RA, Eng P. (2002) GIC effects on pipeline corrosion and corrosion control systems. J. Atmos. Sol. Terr. Phys., 64(16):1755-1764. doi:10.1016/S1364-6826(02)00125-6

[5] Yu Z, Hao J, Liu L, Wang Z. (2019) Monitoring Experiment of Electromagnetic Interference Effects Caused by Geomagnetic Storms on Buried Pipelines in China. IEEE Access, 7:14603-14610. doi:10.1109/ACCESS.2019.2893963

[6] Fernandes De Moraes J, Paulino I, Alves LR, Marcos Denardini C. (2020) Evaluation of possible corrosion enhancement due to telluric currents: Case study of the Bolivia-Brazil pipeline. Ann. Geophys., 38(4):881-888. doi:10.5194/angeo-38-881-2020

[7] Sani MM, Kasuan N, Jusoh MH, Mohd Yassin AI, Hussain Z, Hairuddin MA. (2023) Preliminary Study on the Impact of GIC on Pipe-to-Soil Potential of Buried Pipeline Near Equatorial Region. In Proceedings of the IEEE Symposium on Industrial Electronics and Applications: 15-16 July 2023; Malaysia; pp 1-6. doi:10.1109/ISIEA58478.2023.10212318

[8] Boteler DH. (2013) A new versatile method for modelling geomagnetic induction in pipelines. Geophys. J. Int., 193(1):98-109. doi:10.1093/gji/ggs113

[9] Viljanen A, Pulkkinen A, Pirjola R, Pajunpää K, Posio P, Koistinen A. (2006) Recordings of geomagnetically induced currents and a nowcasting service of the Finnish natural gas pipeline system. Space Weather, 4(10). doi:10.1029/2006SW000234

[10] Sani MM, Yassin AIM, Kasuan N, Ashar NDK, Hairuddin MA, Jusoh MH. (2025) Prediction of the Pipe-to-Soil Potential Based on GIC for Buried Gas Pipeline Using Machine Learning Approach. Journal of Physics: Conference Series, 2998(1):012021. doi:10.1088/1742-6596/2998/1/012021

[11] Rossouw E, Doorsamy W. (2021) Predictive maintenance framework for cathodic protection systems using data analytics. Energies, 14(18):5805. doi:10.3390/en14185805

[12] Yang Y, Liu JJ, Feng XS, Chen PF, Zhang B. (2023) Prediction of the Transit Time of Coronal Mass Ejections with an Ensemble Machine-learning Method. Astrophys. J. Suppl. Ser., 268(2):69. doi:10.3847/1538-4365/acf218

[13] Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. (2020) Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7(1):70. doi:10.1186/s40537-020-00349-y

[14] Dzakyprasetyo, Ferdian D. (2023) Analysis of external corrosion protection performance on buried gas pipeline using CIPS and DCVG methods. ITM Web of Conferences, 61:01022. doi:10.1051/itmconf/20246101022

[15] Patel S, Chamadia PK. (2025) Effect of Solar and Interplanetary Disturbances on Geomagnetic Field: A Comprehensive Review. Journal of Advances in Science and Technology, 22(2):583-596. doi:10.29070/426v7408

[16] Zainuddin A, Hairuddin MA, Yoshikawa A, Helmy M, Abd Latiff ZI, Yassin I, Jusoh H. (2024) Comparative Analysis of dB/dt Response Towards dH/dt as Proxies for Geomagnetically Induced Currents (GICs) in Low-Latitude Region. In Proceedings of the IEEE Symposium on Wireless Technology and Applications: 20-21 July 2024; pp 235-240. doi:10.1109/ISWTA62130.2024.10651826

[17] Viljanen A, Wintoft P, Wik M. (2015) Regional estimation of geomagnetically induced currents based on the local magnetic or electric field. Journal of Space Weather and Space Climate, 5:A24. doi:10.1051/swsc/2015022

[18] Fujimoto A, Uozumi T, Abe S, Matsushita H, Ishitsuka JK, Yoshikawa A. (2016) Long-term EEJ variations by using the improved EE-index. Sun and Geosphere, 11(1):37-47.

[19] Wahyuddin EP, Caraka RE, Kurniawan R, Caesarendra W, Gio PU, Pardamean B. (2025) Improved LSTM hyperparameters alongside sentiment walk-forward validation for time series prediction. Journal of Open Innovation: Technology, Market, and Complexity, 11(1). doi:10.1016/j.joitmc.2024.100458

[20] Chukwura Obi J. (2023) A comparative study of several classification metrics and their performances on data. World Journal of Advanced Engineering Technology and Sciences, 8(1):308-314. doi:10.30574/wjaets.2023.8.1.0054

[21] Sheng L, Ruan X, Wang Y. (2025) Interpretable XGBoost-SHAP machine learning model for identifying scientific breakthroughs. Scientometrics, 130:6801-6832. doi:10.1007/s11192-025-05497-7

[22] Ingham M, Divett T, Rodger CJ, Sigley M. (2025) Observed Effects of the May 2024 Gannon Storm on the New Zealand Gas Pipeline Network-Toward Predicting the Effects of an Extreme Storm. Space Weather, 23(5). doi:10.1029/2024SW004258

[23] Uozumi T, Yumoto K, Kitamura K, et al. (2008) A new index to monitor temporal and long-term variations of the equatorial electrojet by MAGDAS/CPMN real-time data: EE-index. Earth, Planets and Space, 60(7):785-790. doi:10.1186/BF03352828

[24] Caraballo R, González-Esparza JA, Pacheco CR, Corona-Romero P, Arzate-Flores JA, Castellanos-Velazco CI. (2025) The Impact of Geomagnetically Induced Currents (GIC) on the Mexican Power Grid: Numerical Modeling and Observations From the 10 May 2024, Geomagnetic Storm. Geophys. Res. Lett., 52(4). doi:10.1029/2024GL112749

[25] Thaker P, Desai A, Panchal P, Sutaria D, Shah M, Prajapati M. (2025) Evaluating AI approaches for space weather prediction: Strengthening satellite resilience and technology systems. Space Habitation, 1(3).doi:10.1016/j.spaceh.2025.100032