Machine Learning-Augmented Tidal Signature Extraction from GNSS Interferometric Reflectometry (GNSS-IR) Observations

Zulkurnain, I. S.,[1] Sulaiman, S. A.,[1*] Idris, A. H.,[2] Razali, M. H.,[1] Lau, C. L.,[1] Satirapod, C.[3] and Kuo, C.[4]

[1]Surveying Science and Geomatics, Faculty of Alam Bina, Universiti Teknologi MARA, Malaysia
E-mail: izzatsyameer99@gmail.com, saifulaman@uitm.edu.my*
[2]Department of Survey and Mapping Malaysia, Malaysia
[3]Department of Survey Engineering, Chulalongkorn University, Thailand
[4]Department of Geomatics, College of Engineering, National Cheng Kun University, Taiwan
Corresponding Author

DOI: https://doi.org/10.52939/ijg.v22i1.4725

Abstract

GNSS Interferometric Reflectometry (GNSS-IR) is a promising method for sea level monitoring that utilizes reflected GNSS signals to estimate water surface height. It serves as a low-cost and passive alternative to traditional tide gauges, especially beneficial in remote or under-monitored areas. However, conventional spectral techniques used to extract tidal signals from GNSS-IR data often struggle with multipath interference, environmental noise, and rigid model assumptions. To address these limitations, this study proposes a machine learning-based framework that applies three models Random Forest Regression (RFR), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) to enhance the accuracy of tidal height estimation using GNSS signal-to-noise ratio (SNR) data. The approach was tested using data from a GNSS station in Port Klang, Malaysia, with tide gauge measurements serving as reference values. Among the models tested, RFR achieved the best performance with a root mean square error (RMSE) of 0.6451 meters and a coefficient of determination (R^2) of 0.7949. This indicates a strong correlation between the predicted and actual tide data. The findings demonstrate that machine learning techniques, particularly RFR, offer significant improvements over traditional methods and can play a crucial role in enhancing GNSS-IR-based sea level monitoring in regions with limited observational infrastructure.

Keywords: GNSS-IR, Long Short-Term Memory (LSTM), Machines Learning, Random Forest Regression (RFR), Support Vector Regression (SVR).

1. Introduction

Monitoring sea-level variations is vital for coastal hazard mitigation, infrastructure planning, maritime navigation, and understanding the long-term impacts of climate change, including global sea-level rise. Conventional tide gauges, while accurate, are limited by high maintenance costs and sparse spatial distribution, particularly in developing regions and remote coastal zones. Global Navigation Satellite System Interferometric Reflectometry (GNSS-IR) offers a low-cost, passive, and geodetically robust alternative. This technique exploits the interference between direct and surface-reflected GNSS signals captured by a single antenna, most notably observed in oscillations within the signal-to-noise ratio (SNR). These oscillations contain information related to the height of the reflecting surface, enabling the estimation of sea level without additional instrumentation [1].

GNSS-IR has demonstrated utility in several environmental applications, including snow depth retrieval, soil moisture monitoring, and ocean altimetry. However, its accuracy in tidal height estimation remains sensitive to environmental noise, satellite geometry, and local antenna configurations. Traditional methods for extracting tidal signatures from GNSS-IR data, such as sinusoidal curve fitting or the Lomb-Scargle Periodogram, rely on linear or stationary assumptions that do not capture the full complexity of multipath environments and temporal signal variations. Recent advancements in machine learning (ML) offer a powerful solution to these limitations. ML models such as Random Forest

Regression (RFR), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) networks are well-suited for modeling nonlinear and temporal relationships in noisy geodetic datasets. These methods have the potential to extract latent tidal features from SNR observations with improved robustness and predictive performance, even under suboptimal observational conditions.

This study proposes a hybrid GNSS-IR and machine learning framework for the extraction of tidal signatures from GNSS SNR data. Using a case study at Port Klang, Malaysia where GNSS and tide gauge data are co-located we evaluate the performance of RFR, SVR, and LSTM models in predicting daily tidal heights. Our results demonstrate that ML-enhanced GNSS-IR systems significantly outperform traditional spectral methods, providing a scalable, cost-effective approach for real-time and autonomous coastal sea-level monitoring.

2. GNSS-IR Fundamental

GNSS-IR utilizes the interference between direct and reflected satellite signals received by a GNSS antenna. These interactions produce oscillations in the SNR data, the frequency of which is inversely related to the vertical distance between the antenna and the reflecting surface (i.e., sea level). The fundamental relationship is given in Equation 1:

$$h = \frac{\lambda}{2f}$$

Equation 1

Where h is reflector height, λ is the signal wavelength, and f is the frequency of oscillation.

Reflector height can be estimated from SNR oscillations using frequency-domain analysis, but such methods are vulnerable to noise, antenna geometry, and environmental conditions.

The application of GNSS interferometric reflectometry (GNSS-IR) was initially introduced for estimating snow depth and was later extended to monitor coastal sea level changes using reflected GNSS signals. Early studies demonstrated that signal-to-noise ratio (SNR) oscillations, when properly detrended and filtered, contain precise information related to relative reflector height. The development of open-source tools for GNSS-IR data processing further facilitated the broader adoption of this method within the geodetic community, improving its accessibility and reproducibility [2].

Subsequent studies emphasized the value of GNSS-IR for ocean monitoring by utilizing engineered reflectometry features for wave height prediction [3]. The method was also shown to be effective for snow depth retrieval over complex terrains through the use of multi-GNSS constellations, highlighting its versatility across diverse environmental conditions [4]. In addition, recent research has applied deep learning techniques to GNSS-IR observations for soil moisture estimation, reflecting the increasing integration of data-driven approaches within reflectometry studies [5]. While most studies agree on the strong potential of GNSS-IR, several challenges remain, including reduced signal quality at high satellite elevation angles, interference from vegetation or urban environments, and temporal inconsistencies in satellite geometry. These limitations motivate continued improvements in signal preprocessing strategies and encourage the incorporation of machine learning techniques to enhance signal interpretation and prediction accuracy.

3. Materials and Methods

This study utilized Global Navigation Satellite System (GNSS) data from a coastal continuously operating reference station (CORS) located in Port Klang, Malaysia. The observation period spanned two months from June to July 2022, capturing daily signal variations associated with tidal fluctuations. The GNSS data were obtained in Receiver Independent Exchange (RINEX) format and included signal-to-noise ratio (SNR) measurements essential for GNSS Interferometric Reflectometry (GNSS-IR) analysis. GNSS-IR exploits reflected signal components to infer changes in the vertical distance between the antenna and the sea surface, enabling passive and continuous sea-level monitoring.

To ensure the validity of the comparative analysis, both GNSS and tide gauge datasets were carefully aligned in space and time. The geographic coordinates of the GNSS station are approximately 3.05085° N latitude and 101.35638° E longitude. This co-location and temporal consistency between datasets minimize bias due to spatial mismatches and facilitate robust model training and evaluation. An overview of the site configuration and surrounding environment is presented in Figure 1.

Figure 1: GNSS-IR location and site view

To ensure the accuracy and reliability of tidal height estimation from GNSS SNR observations, a comprehensive preprocessing strategy was employed to enhance signal quality and suppress noise. The key steps are as follows:

- Elevation Angle Filtering: GNSS signals with elevation angles between 5° and 15° were retained to capture optimal multipath interference [4]. This range balances signal strength and surface reflection probability.
- Polynomial Detrending: Second- or third-order polynomials were fitted to each SNR track to eliminate low-frequency trends from direct signals. The residuals preserved the multipath-induced oscillations used in RH estimation [6].
- Frequency Analysis via FFT: Fast Fourier Transform (FFT) was used to extract dominant frequencies from the detrended signal, which correspond to the reflector geometry [7].
- Noise and Outlier Filtering: Z-score filtering removed outliers, and an amplitude threshold excluded weak or unstable multipath signals, improving signal reliability.
- Daily Segmentation: Detrended, filtered data were segmented into daily intervals aligned with tide gauge records. Each day was treated as an independent observation unit for model training and validation.

To enhance the extraction and prediction of tidal signatures from GNSS Interferometric Reflectometry (GNSS-IR) data, three machine learning algorithms were implemented: Random Forest Regression (RFR), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) neural networks. These models were chosen for their demonstrated effectiveness in geophysical time series forecasting, nonlinear regression, and environmental remote sensing applications. Random Forest Regression is an ensemble learning method based on decision tree algorithms. It operates by constructing multiple regression trees during training and aggregating their predictions to improve generalization and reduce overfitting. RFR is particularly robust in handling high-dimensional, noisy, and nonlinear geodetic datasets.

- Library: scikit-learn [8]
- Key Hyperparameters:
  - n_estimators: 300 (number of trees)
  - max_depth: None (unlimited depth)
  - min_samples_leaf: 1
  - random_state: 42

Recent studies confirm its utility in GNSS applications. Random Forest Regression (RFR) has been applied to predict wave heights using GNSS reflectometry data [3], while a similar model structure has been employed for soil moisture estimation based on GNSS-IR observations enhanced through Bayesian optimization techniques [5]. Support Vector Regression, derived from Support Vector Machines, is a kernel-based learning algorithm capable of modeling nonlinear relationships between features and target variables. The Radial Basis Function (RBF) kernel is commonly used in SVR for its ability to transform input data into a higher-dimensional space where a linear regression surface can be applied.

- Library: scikit-learn
- Kernel Function: RBF (Radial Basis Function)
- Key Hyperparameters:
  - C: 10 (penalty term)
  - epsilon: 0.1 (insensitive loss zone)
  - gamma: 'scale' (kernel coefficient)

Support Vector Regression (SVR) has been employed for sea surface height (SSH) retrieval using TDS-1 GNSS-R signals, achieving low mean absolute error and strong generalization capability [9]. Similarly, SVR has demonstrated robust performance in wave height modeling and broader environmental forecasting applications [10]. Long Short-Term Memory is a variant of recurrent neural networks (RNNs) that retains long-term dependencies in sequential data through gated memory cells. It is especially effective in forecasting tasks involving diurnal and semi-diurnal oscillations, such as tidal patterns.

- Library: TensorFlow and Keras
- Network Architecture:
  - Input layer: time-lagged sequences of GNSS-IR features (RH, FFT peak frequency, elevation, azimuth)
  - LSTM layer: 64 units
  - Dropout: 0.2
  - Dense output layer: 1 neuron (linear activation)
- Training Parameters:
  - Optimizer: Adam
  - Loss function: Mean Squared Error (MSE)
  - Epochs: 100 (with early stopping)
  - Batch size: 16

Although LSTM has shown strong performance in prior work such as DeepGNSSIR by [11] and hybrid LSTM models for tidal forecasting [12] its application in this study was limited by data availability and potential sequence misconfiguration, resulting in inferior prediction performance relative to RFR and SVR. Each model was trained using feature vectors derived from pre-processed GNSS-IR data, including spectral frequency peaks (via FFT), satellite azimuth and elevation, SNR amplitude, and time-of-day indicators. These features represent the spatiotemporal multipath dynamics and were paired with tide gauge observations to train the predictive models. To ensure robust model performance and generalizability, the dataset was partitioned into training (70%) and testing (30%) subsets. The input features included:

- Daily reflector height (RH) estimates from GNSS-IR preprocessing,
- Satellite azimuth and elevation angles,
- Dominant frequency components extracted from Fast Fourier Transform (FFT),
- Temporal descriptors, such as time-of-day and PRN metadata.

The target variable for supervised learning was the daily mean sea level, measured by co-located tide gauge observations. This setup supports direct validation of GNSS-IR-based estimates against reliable in-situ data.

3.1 Evaluation Metrics

Model performance was assessed using three standard regression metrics:

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)}$$

Equation 2

In Equation 2, *RMSE* denotes the Root Mean Square Error, n represents the total number of observations, and i is the index of the i-th observation. The variable y_i refers to the observed (ground truth) value, while $\hat{y}_i$ denotes the corresponding value predicted by the model. The squared difference $(y_i - \hat{y_i})^2$ measures the magnitude of the prediction error for each observation. RMSE penalizes large prediction errors and reflects the magnitude of deviations. RMSE is one of the most common metrics to evaluate prediction accuracy in regression, time series, and many modeling fields. It measures how far, on average, model predictions are from observed values, in the same units as the target [13][14] and [15].

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Equation 3

In Equation 3, *MAE* represents the Mean Absolute Error, n denotes the number of observations, and i is the observation index. The term y_i corresponds to the observed value, whereas $\hat{y}_i$ indicates the predicted value. The absolute value operator $|y_i - \hat{y_i}|$ computes the absolute difference between the observed and predicted values for each data point. MAE provides a linear penalty for all errors and is more robust to outliers. Mean Absolute Error (MAE) is a basic but powerful way to measure how far model predictions are, on average, from true values in regression and forecasting. It is widely used in machine learning, climate science, signal processing, and many applied fields [16] and [17].

- Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Equation 4

In Equation 4, R^2 denotes the Coefficient of Determination, y_i represents the observed value, $\hat{y}_i$ is

the predicted value, and $\bar{y}$ indicates the mean of all observed values. R^2 measures how well the model explains the variance in the ground truth data.
These metrics provide complementary insights into both absolute prediction accuracy and variance explanation, in line with best practices in geodetic and environmental regression modeling [9].

3.2 Model Optimization and Validation Strategy
Each machine learning model was optimized using tailored validation strategies:

- Random Forest Regression (RFR) and Support Vector Regression (SVR) were tuned using 5-fold cross-validation, implemented via GridSearchCV in scikit-learn [8]. Hyperparameters such as the number of trees (n_estimators) for RFR and the penalty coefficient (C) and kernel parameter (gamma) for SVR were adjusted to minimize cross-validated RMSE.
- Long Short-Term Memory (LSTM) networks were validated using a walk-forward validation strategy. This approach preserves the chronological order of the time series by training on earlier time windows and testing on the next time step—simulating a realistic forecasting environment [12].
- Early stopping and dropout regularization were applied to LSTM to prevent overfitting, using Keras callbacks such as EarlyStopping(monitor='val_loss'). Input data were reshaped into 3D tensors [samples, time_steps, features] to conform with sequential learning requirements [11].

Data normalization was applied using StandardScaler() from scikit-learn, and the same transformation was applied consistently to training and test sets. Feature engineering was informed by domain knowledge from recent GNSS-IR publications [5].

4. Results and Discussions
4.1 Reflection Zone
To confirm that GNSS signals reflect off the sea surface, this study employed the First Fresnel Zone (FFZ) as a geometric criterion for identifying viable reflection areas. The FFZ delineates the region around the specular reflection point where constructive and destructive interference can significantly affect the received signal. Using precise satellite positions from IGS orbit products (*.sp3) and receiver coordinates from the RINEX header, the azimuth and elevation angles of GNSS satellites were computed for every observation epoch. The FFZ was modeled as an elliptical footprint on the local horizontal plane. The size and shape of this ellipse vary with:

Satellite elevation angle (lower angles produce wider zones),
- Reflector height, and
- Signal frequency (e.g., L1).

As the satellite elevation increases, the FFZ becomes smaller and moves closer to the GNSS antenna, reducing the strength of the reflected signal. This behavior is consistent with theoretical and observational findings in GNSS-IR literature [4].

Figure 2 visualizes the spatial extent of the FFZ assuming a reflector height of 5 meters. This analysis confirms that sea-surface reflections dominate the GNSS-IR signature at low elevation angles, validating the coastal positioning of the Port Klang GNSS station for tidal monitoring. The impact of satellite elevation angle on GNSS-IR signal quality was evaluated by applying three elevation masks: 0°–5°, 5°–10°, and 10°–15°. This stratification is commonly adopted in GNSS-IR studies to optimize multipath sensitivity while minimizing atmospheric and structural noise [3] and [7]. Figure 2 shows an illustration related to the Fresnel zone at the station based on azimuth angle. The size of the Fresnel zone ellipse is related to the reflector height, the satellite elevation angle, and the signal frequency used.

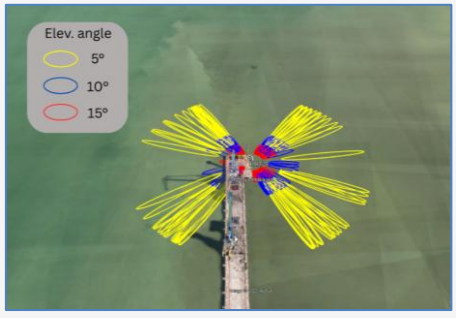

Figure 2: Reflection zone at station plotted by assuming reflector height is 5 meters

4.2 Plot SNR Periodogram for Elevation Angle: 0°–5°, 5°–10°, and 10°–15° *(DOY153)*
Figure 3 presents the signal-to-noise ratio (SNR) periodograms for three elevation angle ranges: 0°–5°, 5°–10°, and 10°–15°, corresponding to Day of Year (DOY) 153. These elevation bins were selected to evaluate how satellite elevation angle influences multipath signal strength, spectral characteristics, and data reliability in GNSS-IR tidal analysis. The 0°–5° elevation range (Figure 3(a)) exhibits the strongest multipath signals due to the extremely shallow incidence angle of GNSS signals interacting with the sea surface. At such low elevation angles,

reflected signals travel longer paths before reaching the antenna, resulting in strong constructive and destructive interference patterns. This is clearly reflected in the SNR periodogram, which shows high spectral power and pronounced periodic oscillations, indicating strong sensitivity to reflector height variations. However, despite its high multipath strength, this elevation range is also the most vulnerable to several error sources. These include signal obstruction by nearby buildings, terrain irregularities, or vegetation, as well as increased sensitivity to tropospheric delay errors due to the longer signal path through the atmosphere. In addition, antenna gain pattern distortions are more pronounced at very low elevation angles, potentially introducing biases into the observed SNR oscillations. The 5°–10° elevation range (Figure 3(b)) represents a balanced trade-off between multipath signal strength and observational stability. In this range, the reflected GNSS signals remain sufficiently strong to produce clear SNR oscillations, while the influence of obstructions and atmospheric disturbances is significantly reduced compared to the lowest elevation bin. The SNR periodogram for this range displays distinct and well-defined spectral peaks, indicating cleaner and more stable periodic behavior. This suggests that the 5°–10° elevation range provides optimal conditions for GNSS-IR analysis, offering reliable reflector height sensitivity with reduced noise contamination. As a result, this elevation range is often preferred in practical GNSS-IR applications for sea-level and tidal monitoring.

At higher elevation angles, specifically 10°–15° (Figure 3(c)), the GNSS signals experience substantially reduced multipath interference. As the elevation angle increases, the specular reflection point moves closer to the antenna and gradually exits the first Fresnel zone (FFZ). Consequently, the SNR observations in this range show lower spectral power and fewer pronounced oscillations, leading to a cleaner but less multipath-sensitive signal. While this results in noise-free SNR measurements, the reduced multipath content limits the effectiveness of this elevation range for precise reflector height estimation. Therefore, higher elevation angles are generally less suitable for GNSS-IR tidal analysis, although they remain valuable for quality control and validation purposes.

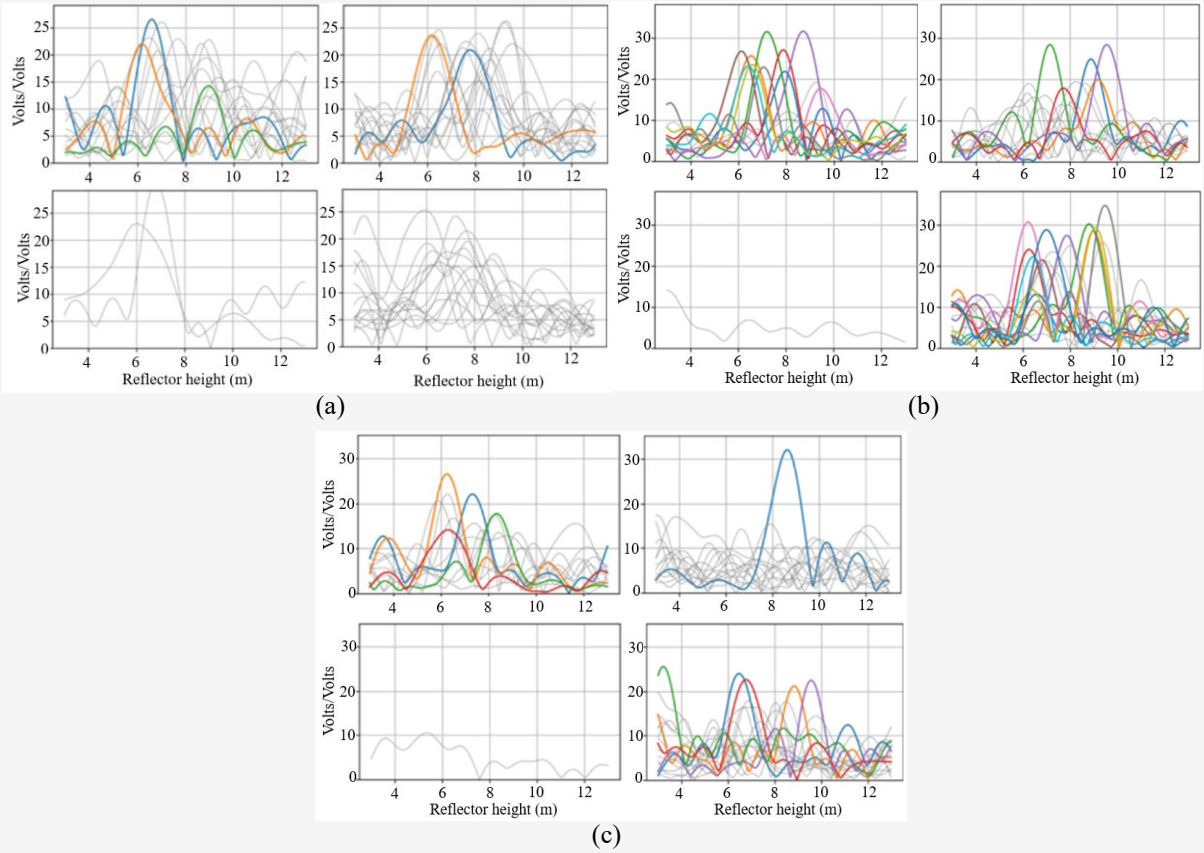

Figure 3: Plot SNR periodogram for elevation angle: (a) 0°–5°, (b) 5°–10°, and (c) 10°–15°

4.3 Plot of Selected Reflected Height for Elevation Angle: 0°–5°, 5°–10°, and10°–15° (DOY153)

Based on Figure 4(a), although strong multipath effects are clearly present at low elevation angles, these signals also introduce noticeable noise and irregular scatter in the reflector height (RH) estimates. The RH values show larger variability across azimuth directions, with a higher proportion of observations failing the quality control thresholds, as indicated by the scattered points below the peak-to-noise and spectral amplitude criteria. This behavior is consistent with previously reported cautionary findings [6], where low-elevation GNSS reflections were shown to be highly sensitive to environmental obstructions, antenna gain effects, and atmospheric disturbances. Low-elevation GNSS signals travel through a longer atmospheric path and are more affected by tropospheric and ionospheric refraction and scintillation, leading to additional delay and signal fading.

In contrast, Figure 4(b) demonstrates that RH estimates within the 5°–10° elevation range are more stable and consistently distributed across azimuth angles. A larger number of observations pass the quality control thresholds for peak-to-noise ratio and spectral peak amplitude, indicating stronger and more reliable multipath signatures. The RH values in this range show improved coherence and align more closely with expected tidal trends, suggesting that this elevation range offers a favorable balance between multipath strength and observational stability. At higher elevation angles (10°–15°), as shown in Figure 4(c), the oscillation amplitudes are noticeably reduced, as reflected by lower spectral peak amplitudes and fewer strong peaks passing the quality control criteria. While the SNR observations appear cleaner with reduced noise, the weakened multipath signal results in diminished sensitivity to reflector height variations. Consequently, RH estimates in this range are less responsive to tidal fluctuations. Although such conditions may be useful in environments with high sky visibility, this elevation range may underperform in capturing small-scale tidal variability, a pattern that is consistent with recent findings from DeepGNSSIR experiments [11].

Overall, the 5°–10° elevation range provides the most reliable and consistent performance, effectively balancing multipath richness with environmental stability. This observation agrees with recommendations from other recent GNSS-IR studies focused on oceanographic monitoring [3] and [4].

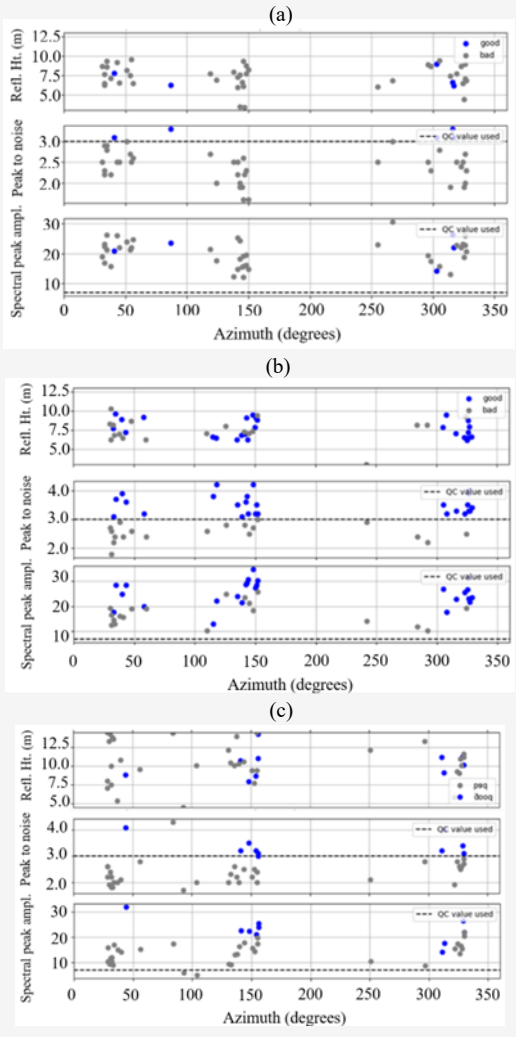

Figure 4: Plot of selected reflected height for elevation angle: (a) 0°–5°, (b) 5°–10°, and (c) 10°–15°

4.4 The Random Forest Regression (RFR)

The Random Forest Regression (RFR) model exhibited the most robust performance in predicting reflector height (RH) from GNSS-IR observations. Its ensemble architecture, which aggregates predictions from multiple decision trees, enables it to effectively capture nonlinear interactions among geophysical input features, making it particularly well-suited for complex and noisy GNSS-IR data. Figure 5 presents the model's feature importance rankings. Among all variables, satellite azimuth angle (Azim) emerged as the most influential. This is consistent with theoretical expectations, as the azimuth angle governs the direction of the satellite's motion across the sky and significantly impacts the geometry of the reflection zone and interference patterns in the SNR signal [4] and [3].

The prominent role of azimuth has also been confirmed in GNSS-IR feature engineering studies where it improves signal discrimination over water bodies. The Random Forest Regression model uses several GNSS-IR-derived features to predict reflector height and tidal variations. Azim (Azimuth) represents the satellite's horizontal direction from north and strongly affects reflection geometry. UTCtime captures temporal variations like tidal phase and satellite geometry. Amp (Spectral Amplitude) indicates the strength of surface reflections, while EdotF (Elevation Rate Factor) reflects how quickly the satellite elevation changes, influencing signal patterns. PkNoise (Peak Noise Ratio) distinguishes clean from noisy signals, and DelT (Time Duration) measures the satellite track length, affecting frequency resolution. Other features include freq (Frequency), which relates directly to reflector height, and eminO/emaxO (minimum/maximum elevation angles), which modulate reflection sensitivity and noise. Rise indicates whether the satellite is rising or setting, with minimal influence.

UTC time contributed significantly to prediction accuracy, suggesting that temporal periodicities such as diurnal or tidal cycles: are encoded in the signal behavior and are critical for RH estimation. This observation aligns with findings in [11], who incorporated temporal metadata to improve GNSS-IR tidal prediction in DeepGNSSIR. Moderate contributions were noted from features such as:

- SNR amplitude (Amp) – indicating signal strength variations,
- Elevation rate of change (EdotF) – which reflects satellite motion relative to the horizon,

Peak noise (PkNoise) and delay time (DelT) – likely capturing signal variability near the Fresnel boundary

In contrast, features such as eminO, emaxO, and rise had lower importance scores, which may reflect redundancy or weak correlation with RH for this site and time period. These findings highlight the interpretability advantage of RFR over black-box models like LSTM, allowing researchers to better understand the physical relevance of each input. Figure 6 illustrates the predicted versus actual RH values using RFR. The close alignment along the 1:1 diagonal line indicates high predictive accuracy and minimal bias. The model effectively captures both low and high RH values without introducing significant outliers, demonstrating excellent generalization across diverse signal conditions. This pattern is further supported in Figure 7, where the RFR-predicted tide heights are plotted alongside tide gauge ground truth. The RFR model not only tracks the amplitude of the tidal cycle but also preserves its temporal phase, successfully capturing peaks and troughs over the entire two-month observation period. Such performance is consistent with previous studies reporting that Random Forest Regression (RFR) models outperform neural networks and linear regression approaches in tidal signal prediction using GNSS-derived features [12]. Similarly, RFR combined with appropriate feature selection strategies has been shown to achieve high accuracy in GNSS-IR applications, including soil moisture and sea level monitoring [5]. In summary, the RFR model's ability to manage noisy, nonlinear, and multicollinear data, alongside its interpretability and minimal parameter tuning requirements, makes it a strong candidate for operational deployment in coastal GNSS-IR tide monitoring systems.

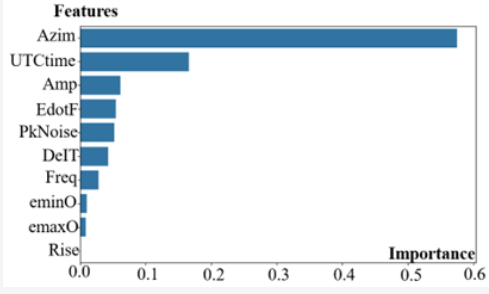

Figure 5: Random Forest feature importance

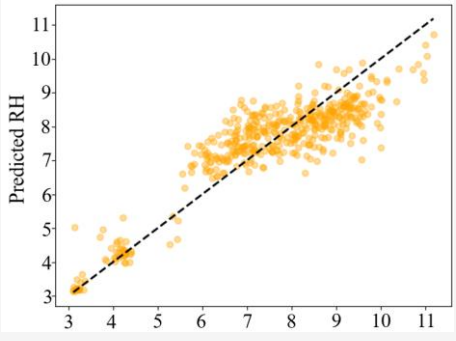

Figure 6: Predicted vs Actual RH (RFR)

4.5 The Support Vector Regression (SVR)

The Support Vector Regression (SVR) model demonstrated moderate predictive performance in estimating reflector height (RH) from GNSS-IR-derived features. As illustrated in Figure 8, the SVR-predicted RH values (blue scatter points) exhibit a general positive correlation with the observed values, indicating the model captured some meaningful relationships within the data. However, the predictions showed greater dispersion around the 1:1 reference line compared to Random Forest Regression (RFR), particularly in the mid-to-high RH range.

This wider spread suggests the model may be affected by either underfitting or overfitting, likely stemming from suboptimal hyperparameter tuning a known challenge in SVR applications. Key parameters such as the penalty term (C), kernel coefficient (γ), and epsilon-insensitive loss margin (ε) strongly influence the regression margin and flexibility of the model [9] and [10]. An improperly tuned SVR can either fail to capture complex data variations or overreact to noisy patterns, both of which appear in this case. GNSS-IR data is inherently noisy and nonlinear, and while SVR is theoretically capable of modeling such complexity via its kernel functions (especially RBF), its performance often declines in high-dimensional or low-signal-to-noise settings, unless supported by rigorous feature selection and normalization [13]. Nonetheless, the underlying trend between predicted and observed reflector height (RH) remains discernible, indicating that Support Vector Regression (SVR) is capable of learning a fundamental mapping between input features and tidal response. With improved preprocessing strategies and automated hyperparameter optimization techniques such as Bayesian optimization or genetic algorithms, SVR has the potential to narrow the performance gap relative to Random Forest Regression (RFR), as demonstrated in previous GNSS-IR soil moisture studies [5]. Figure 9 further assesses SVR performance over time by comparing predicted RH values to co-located tide gauge records across the two-month dataset.

The SVR predictions follow the general oscillatory pattern, but they appear overly smoothed and lack sensitivity to rapid changes in tidal amplitude. This low temporal resolution may limit SVR's utility in dynamic coastal environments where short-term fluctuations are critical. This smoothing effect has been observed in other geophysical applications where SVR favors overall trend preservation over localized variability [12]. While this conservative behavior reduces susceptibility to outliers, it can compromise the timing and peak amplitude accuracy required for real-time sea level monitoring or warning systems. Overall, the SVR model shows reasonable but limited performance:

- It captures general trends in GNSS-IR data and aligns with mean sea-level behavior.
- It underperforms in capturing extreme RH variations, especially compared to ensemble methods like RFR.
- It is sensitive to data scaling, kernel choice, and hyperparameter tuning, requiring extensive optimization to be competitive.

With further refinement particularly in feature selection, scaling, and automated parameter search; SVR remains a viable alternative for GNSS-IR-based sea level modeling, especially in settings with moderate variability and limited training data.

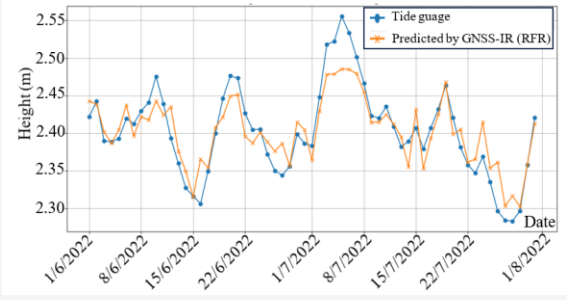

Figure 7: Tide Gauge vs GNSS-IR (RFR)

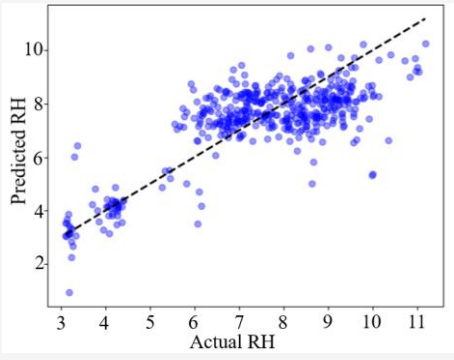

Figure 8: Actual vs Predicted RH (SVR)

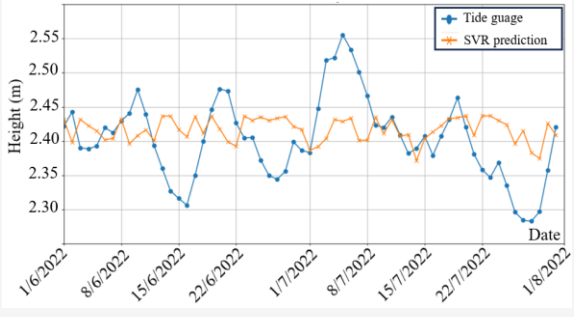

Figure 9: GNSS-IR vs Tide Gauge (SVR)

4.6 Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) model, a type of recurrent neural network (RNN), was implemented to capture temporal dependencies in GNSS-IR-derived reflector height (RH) sequences. LSTM is a special type of recurrent neural network (RNN) designed to learn from sequential data (time series, text, speech) by remembering information for long periods and avoiding the classic vanishing gradient problem of standard RNNs [18][19] and [20].

However, as shown in Figure 10, the model underperformed, producing nearly constant predictions clustered around 7.5 meters regardless of the actual RH values. This flat prediction behavior typically indicates premature convergence or mode collapse, where the model defaults to outputting the mean of the training data. Such outcomes are common when:

- Input sequences are not formatted correctly (i.e., missing 3D structure: [samples, time_steps, features]),
- The dataset lacks sufficient temporal resolution or variability,
- The model architecture is too shallow or overly regularized (e.g., via dropout), or
- Training configurations (e.g., learning rate, batch size, optimizer) are poorly tuned [11] and [12].

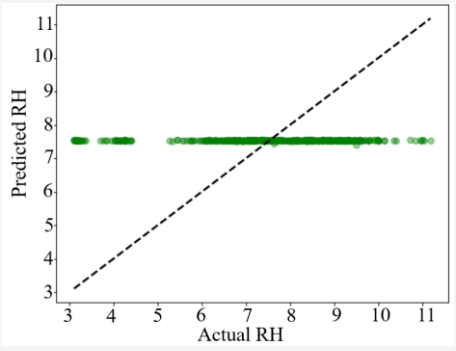

Figure 10: Actual vs Predicted RH (LSTM)

LSTM models are designed to retain memory of long-term sequential patterns through gating mechanisms. However, they are data-hungry and computationally sensitive, especially in domains like GNSS-IR where the temporal signature of the signal is sparse, noisy, or irregular [4]. Without robust temporal continuity or sufficient training volume, the model may fail to extract meaningful patterns precisely what appears to have occurred here. Figure 11 compares LSTM-predicted RH values against actual tide gauge observations for the July 15–31 period. While the prediction curve broadly captures the directional trends of tidal behavior (i.e., rising and falling phases), it consistently overestimates the amplitude and lacks the granularity to replicate short-term oscillations. This smoothing behavior is a known limitation of unoptimized LSTM models, particularly when trained on limited or improperly segmented time series [3] and [5].

Despite the theoretical advantages of LSTM in modeling nonlinear geophysical time series, this study illustrates that deep learning is not inherently superior without:

- Proper temporal segmentation and windowing strategies,
- Sufficient training epochs and data volume,
- Careful architecture design (e.g., stacked layers, bidirectional LSTM), and
- Thoughtful hyperparameter tuning

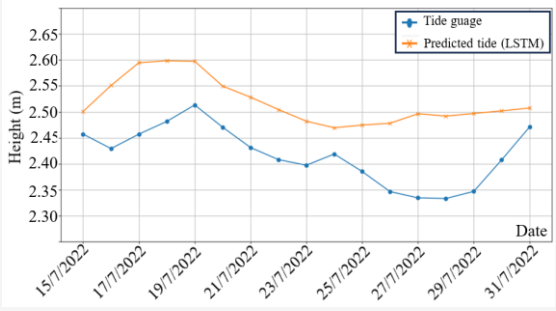

Figure 11: GNSS-IR vs Tide Gauge (LSTM)

Table 1: Performance metrics summary

Model	MSE (m²)	R^2
RFR	0.645	0.79
SVR	1.287	0.59
LSTM	3.105	0.00

These challenges are echoed in other GNSS-IR and GNSS-R studies that report LSTM underperformance when facing limited training sets or irregular observation schedules [11] and [12]. While LSTM models offer a promising approach for sequence-based GNSS-IR tidal estimation, their success hinges on:

- Well-structured temporal data,
- Adequate training volume, and
- Thoughtful model design and optimization.

In this study, the LSTM model failed to meet these criteria and was thus unsuitable for RH regression without major reconfiguration. Future work may benefit from using hybrid models (e.g., CNN-LSTM, attention-based LSTM) or incorporating data augmentation techniques to improve temporal learning in sparse GNSS-IR datasets.

4.7 Performance Metric Summary

The quantitative evaluation of the three machine learning models RFR, SVR, and LSTM are summarized in Table 1. RFR achieved the lowest MSE (0.645 m²) and the highest R^2 score (0.79), indicating excellent predictive performance and a strong model fit. An R^2 value of ~0.80 suggests that nearly 80% of the variance in the observed reflector height (RH) was explained by the model. This result aligns with the earlier visual assessments (Figures 7 and 8), where RFR predictions closely tracked both

RH and tide gauge observations across the full range of tidal variation.

The strength of RFR lies in its ensemble structure, which reduces variance by averaging predictions across multiple decision trees. This makes it highly effective for noisy and nonlinear datasets, such as those encountered in GNSS-IR sea-level sensing [3] and [5]. The model's resilience to multicollinearity and missing data further enhances its reliability in real-world coastal monitoring scenarios. SVR demonstrated moderate performance, with an MSE of 1.287 m² and an R^2 of 0.59. The R^2 value suggests that SVR explained approximately 59% of the variance in the RH observations a meaningful but significantly weaker fit than RFR. These results reinforce known characteristics of SVR: while capable of modeling nonlinear trends, its performance is highly sensitive to kernel selection and hyperparameter tuning. The use of a radial basis function (RBF) kernel provides flexibility, but improper tuning of the penalty term (C), kernel scale (γ), and error margin (ε) can lead to either underfitting or overfitting [9] and [10]. SVR remains a promising option for GNSS-IR applications, especially when combined with advanced optimization strategies like Bayesian search or cross-validation.

The LSTM model produced the poorest performance, with the highest MSE (3.105 m²) and a negative R^2 value. The negative R^2 indicates that the model performed worse than a naïve baseline predictor that simply outputs the mean of the target variable. This result confirms earlier observations (Figure 11), where the LSTM model failed to capture tidal dynamics and instead generated nearly flat output. Possible causes include:

- Incorrect input formatting for sequence modeling (e.g., missing 3D tensor shape: [samples, time_steps, features]),
- Insufficient temporal resolution or data volume,
- Inadequate model depth or hyperparameter tuning,
- Vanishing gradient problems during training.

These issues are well-documented in the GNSS-IR literature, where deep learning models like LSTM often require extensive preprocessing, data augmentation, and architectural experimentation to perform effectively [4] and [5]. The comparative evaluation of MSE and R^2 metrics clearly supports the findings of the visual analysis:

- Random Forest Regression emerges as the most robust and accurate model, offering high prediction fidelity with relatively low tuning effort.
- SVR provides moderate predictive utility, contingent upon hyperparameter optimization.
- LSTM, in its current form, is not suitable for GNSS-IR RH regression due to insufficient data structure and training.

These findings are consistent with recent studies in the field, which indicate that Long Short-Term Memory (LSTM) models tend to outperform classical approaches only when trained on large, well-structured datasets with strong temporal dependencies [11] and [12]. In contrast, ensemble-based methods such as Random Forest Regression (RFR) remain among the most consistent performers in GNSS-IR and GNSS-R applications due to their flexibility, interpretability, and robustness to noisy input features [3] and [7].

5. Conclusions

This study presented a comprehensive investigation into the application of GNSS Interferometric Reflectometry (GNSS-IR) for sea level monitoring, enhanced through the integration of machine learning (ML) techniques. The primary objective was to improve the accuracy and robustness of tidal signal extraction and reflector height (RH) prediction from GNSS SNR data using supervised learning approaches. Using a two-month dataset (June–July 2022) collected from a coastal GNSS station in Port Klang, Malaysia, this research aligned GNSS observations with co-located tide gauge measurements, enabling rigorous performance evaluation. Through an extensive preprocessing pipeline, SNR detrending, spectral analysis, and feature engineering, the study successfully isolated usable multipath signals linked to tidal variations. These refined features served as the foundation for RH estimation and subsequent machine learning model training.

Three supervised ML models were evaluated:

- Random Forest Regression (RFR),
- Support Vector Regression (SVR), and
- Long Short-Term Memory (LSTM) neural networks.

Among them, RFR outperformed the others, achieving an RMSE of 0.645 meters and an R^2 of 0.79. Its ensemble structure enabled it to capture nonlinear dependencies, handle noisy inputs, and provide interpretable insights through feature importance analysis. These strengths make RFR a highly reliable and practical model for GNSS-IR-based sea level estimation, consistent with recent findings in similar geodetic and remote sensing applications [3] and [12]. The SVR model showed moderate predictive capability, with an RMSE of 1.287 m and R^2 of 0.59. Although SVR was able to

approximate overall tidal trends, it was less responsive to short-term variations, a limitation attributed to its sensitivity to kernel and parameter tuning as similarly noted in studies of GNSS-R sea surface estimation.

The LSTM model performed poorly with an RMSE of 3.1051 m and a negative R^2 score. Its failure to extract temporal patterns likely stemmed from insufficient training data, improper sequence formatting, and suboptimal network architecture. This result highlights a critical insight: deep learning models require large, well-structured, and temporally coherent datasets to be effective in GNSS-IR applications [5] and [11]. This study demonstrates that GNSS-IR, when paired with machine learning, offers a scalable and cost-effective alternative to traditional tide gauges, particularly in infrastructure-limited or remote coastal regions. Among the tested models, RFR emerges as the most promising candidate for operational deployment, balancing predictive performance, model interpretability, and data requirements. Looking ahead, the framework could be further improved by:

- Enhancing deep learning architectures (e.g., CNN-LSTM, attention mechanisms),
- Incorporating auxiliary datasets, such as meteorological, oceanographic, or atmospheric delay data,
- Applying the model across diverse coastal sites to assess spatial generalizability and transferability,
- And evaluating the feasibility of near real-time tidal monitoring using multi-constellation GNSS data streams.

By addressing these enhancements, the integration of GNSS-IR with intelligent ML systems has the potential to contribute to next-generation sea level monitoring solutions, supporting both scientific research and coastal resilience planning in the face of rising seas and climate variability.

Acknowledgments

We acknowledge the Department of Survey and Mapping Malaysia (JUPEM) for providing GNSS data. This research was supported by Land Surveyors Boards Malaysia

References

[1] Tabibi, S., Sauveur, R., Guerrier, K., Metayer, G. and Francis, O., (2021). SNR-Based GNSS-R for Coastal Sea-Level Altimetry. *Geosciences,* Vol. 11(391). https://doi.org/10.3390/geosciences11090391.

[2] Roesler, C. and Larson, K. M., (2018). Software Tools for GNSS Interferometric Reflectometry (GNSS-IR). *GPS Solutions,* Vol. 22(3). https://doi.org/10.1007/s10291-018-0744-8.

[3] Becker, J. M. and Roggenbuck, O., (2023). Prediction of Significant Wave Heights with Engineered Features from GNSS Reflectometry. *Remote Sensing,* Vol. 15(3). https://doi.org/10.3390/rs15030822.

[4] Zhou, W., Liu, Y. and Huang, L., (2022). Multiconstellation GNSS Interferometric Reflectometry for the Correction of Long-term Snow Height Retrieval on Sloping Topography. *GPS Solutions,* Vol. 26. https://doi.org/10.1007/s10291-022-01333-0.

[5] Daneghian, P. and Rastbood, A., (2024). GNSS-IR Soil Moisture Estimation Using Deep Learning with Bayesian Optimization for Hyperparameter Tuning. *Journal of Geodetic Science,* Vol. 14(1). https://doi.org/10.1515/jogs-2022-0172.

[6] Martín, A., Luján, R. and Julián, A., B. A., (2020). Python Software Tools for GNSS Interferometric Reflectometry (GNSS-IR). *GPS Solutions,* Vol. 24(4). https://doi.org/10.1007/s10291-020-01010-0.

[7] Wang, N., Xu, T., Gao, F., He, Y., Meng, X., Jing, L. and Ning, B., (2022). Sea-level Monitoring and Ocean Tide Analysis Based on Multipath Reflectometry Using Multi-GNSS Signals. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 60. https://doi.org/10.1109/TGRS.2022.3219074.

[8] Pedregosa, F., Varoquaux, G. and Gramfort, A., (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Vol. 12, 2825–2830.

[9] Zhang, Y., Huang, S. and Han, Y., (2022). Machine Learning Methods for Spaceborne GNSS-R Sea Surface Height Measurement from TDS-1. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 15, 1079–1088. https://doi.org/10.1109/JSTARS.2021.3139376.

[10] Shamshirband, S., Mosavi, A., Rabczuk, T., Nabipour, N. and Chau, K., (2020). Prediction of Significant Wave Height; Comparison Between Nested Grid Numerical Model, and Machine Learning Models. *Engineering Applications of Computational Fluid Mechanics,* Vol. 14(1), 134–149. https://doi.org/10.1080/19942060.2020.1773932.

[11] Chen, L., Zhang, W. and Sun, J., (2023). Deep GNSSIR: Deep Learning Model for GNSS-IR Tidal Prediction. *Remote Sensing of Environment,* Vol. 294. https://doi.org/10.1016/j.rse.2023.113514.

[12] Ban, W., Shen, L., Lu, F., Liu, X. and Pan, Y., (2023). Research on Long-term Tidal-height-prediction-based Decomposition Algorithms and Machine Learning Models. *Remote Sensing,* Vol. 15(12). https://doi.org/10.3390/rs15123045.

[13] Aziz, M., Pa'suya, M., Talib, N., Din, A., Hashim, S., and Ramli, M. (2023). Vertical Accuracy Assessment of Improvised Global Digital Elevation Models (MERIT, NASADEM, EarthEnv) Using GNSS and Airborne IFSAR DEM. *International Journal of Geoinformatics*, Vol. 19(12), 65–82. https://doi.org/10.52939/ijg.v19i12.2979.

[14] Chai, T. and Draxler, R. R., (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature. *Geoscientific Model Development,* Vol. 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

[15] Black, J. L., Conolly, L. W., Flint, K., Grundy, I., Lepan, D., Liuzza, R. M., Mcgann, J. J., Prescott, A. L., Qualls, B. V. and Waters, C., (2014). *The Broadview Anthology of British literature.* Broadview Press.

[16] Qi, J., Du, J., Siniscalchi, S. M., Ma, X. and Lee, C. H., (2020). On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression. *IEEE Signal Processing Letters,* Vol. 27, 1485–1489. https://doi.org/10.1109/lsp.2020.3016837.

[17] Botchkarev, A., (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management,* Vol. 14, 045–076. https://doi.org/10.28945/4184.

[18] Sagheer, A. and Kotb, M., (2019). Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. *Scientific Reports,* Vol. 9(1). https://doi.org/10.1038/s41598-019-55320-6.

[19] Sak, H., Senior, A. and Beaufays Google, F., (2014). *Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition.* https://arxiv.org/pdf/1402.1128.

[20] Karim, F., Majumdar, S., Darabi, H. and Chen, S., (2018). LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access,* Vol. 6, 1662–1669. https://doi.org/10.1109/access.2017.2779939.