Jurnal Teknologi Informasi dan Komunikasi

Volume 16 Issue 2 Year 2025 | Page 135-149 | e-ISSN: 2477-3255 | ISSN: 2086-4884
Received: 31-07-2025 | Revised: 05-09-2025 | Accepted: 10-10-2025

Modelling the Hatching Success of Sea Turtle Eggs Using Long Short-Term Memory (LSTM) for Conservation Oriented Ecotourism

Agustriono[1,5], Susanti[2], Lusiana[3], Mardainis[4], Rahmat Irfansyah[5]
[1,2,3,4] Universitas Sains dan Teknologi Indonesia, Pekanbaru, Indonesia, 28292
[5] Loka Kawasan Konservasi Perairan Nasional Pekanbaru, Pekanbaru, Indonesia, 28286
E-mail: agustriono.neckoe@gmail.com[1,5], susanti@usti.ac.id[2], lusiana@usti.ac.id[3], mardainis@usti.ac.id[4], rahmat.irfansyahku@gmail.com[5].
*Corespondence: agustriono.neckoe@gmail.com

Abstract: *This study proposes a Long Short-Term Memory (LSTM) model to predict the hatching success of sea turtle eggs in the Anambas Islands Marine Conservation Area, Indonesia. Leveraging nesting data (2022–2024) provided by LKKPN Pekanbaru and associated environmental variables, the model's performance was assessed across various configurations of time steps (2, 5, 7, 30, and 45 days) and data splits (ranging from 60:40 to 90:10). The optimal configuration—7-day time step with a 60:40 train-test split—yielded RMSE = 17.90, MAE = 8.67, and R^2 = 0.34. Results revealed strong seasonal nesting trends and statistically significant interspecies differences in incubation periods ($p < 0.05$). While the model demonstrated high predictive accuracy for standard incubation durations (30–45 days), performance declined in extreme cases, highlighting the need for location-specific environmental data. This research illustrates the practical application of LSTM for ecological time series forecasting and provides a machine learning framework to support decision-making in ecotourism scheduling and marine conservation planning in island-based coastal ecosystems.*

Keywords: *Conservation, LSTM, Sea Turtle, Mangkai, MAE, RMSE, Ecoturism, Anambas*

1. Introduction

Indonesia's marine region, particularly the Anambas Islands Regency, possesses substantial potential for marine biodiversity and resource development [1]. The marine ecosystem serves as a critical habitat for sea turtles (Cheloniidae), species that are currently threatened with extinction due to continuous population decline [2]. This decline can be mitigated through controlled hatching practices that ensure the continuity of the turtles' life cycle [3]. The success of sea turtle hatcheries depends largely on the availability and sustainability of healthy marine ecosystems, where well-managed areas play a strategic role in conservation and ecotourism activities, implemented through semi-natural or intensive hatching approaches [4]. Effective conservation of sea turtles requires intensive, collaborative monitoring efforts among various stakeholders, including local communities, conservation institutions, and government agencies [5].

Mangkai Island, located within the Anambas Islands Marine Conservation Area, represents one of the primary nesting sites for sea turtles [6]. Geographically situated at 03°05'32" N and 105°35'00" E and covering approximately 2.27 km², this island is recognized as an ecologically vital zone. Monitoring data collected by LKKPN Pekanbaru from 2022 to 2024 reveal a consistent seasonal nesting pattern, with peak activity occurring between May and September, and the highest frequency recorded in June and July. The total number of nesting events was reported as 2,641 in 2022, 2,851 in 2023, and 2,532 in 2024 [7]–[9]. These consistent temporal trends highlight the importance of systematic natural resource management, ensuring that ecological benefits can be sustainably enjoyed by local communities while maintaining biodiversity integrity

[2]. Species-based conservation management, particularly for sea turtles, aims not only to enhance conservation success but also to improve community livelihoods through sustainable utilization [10]. Furthermore, species conservation offers opportunities for ecotourism development, which can generate alternative income for local communities under conservation-based programs [9].

Promotional activities carried out by LKKPN Pekanbaru have demonstrated an increasing trend in tourist arrivals, with 76 visitors in 2022, 122 in 2023, and 269 in 2024. The majority of visits coincide with the sea turtle nesting season from April to November. Data on tourist activities within the Anambas Islands and the National Marine Conservation Area show that 43.5% of visitors engaged in snorkeling, 24.6% in diving, 3.6% in turtle watching, 2.2% in fishing, 0.7% in survival training, and 25.4% in other recreational activities [11]. The low proportion of turtle-watching tourists, despite the area's ecological potential, underscores the need for enhanced conservation-based ecotourism education and data-driven planning by conservation managers. However, a key limitation remains: no reliable predictive system currently exists to accurately estimate the hatching success rate of sea turtle eggs, which is essential for both conservation decision-making and ecotourism management. As a result, conservation practices remain largely reactive, relying on field observations that are often time-consuming, costly, and prone to environmental variability.

To address this challenge, this study proposes the use of Long Short-Term Memory (LSTM) networks to predict the hatching success rate of sea turtle eggs. The research framework is built upon marine ecological theory, temporal data modeling, and information technology-based systems. LSTM, a subclass of Recurrent Neural Networks (RNNs)[12]. Is specifically designed to recognize and learn sequential dependencies in time-series data [13]. Since sea turtle nesting and hatching cycles exhibit strong seasonal and temporal regularities [14], LSTM provides an appropriate computational model for predicting these ecological phenomena. Its capacity to capture nonlinear and long-term dependencies makes it a promising tool for ecological forecasting and conservation management. The objective of this study is to develop a robust and interpretable prediction model for sea turtle hatching success based on ecological and temporal variables. The proposed model aims to assist conservation managers in designing strategic turtle-watching ecotourism programs, providing accurate information to visitors, and contributing to non-tax state revenue (PNBP) through sustainable marine-based ecotourism initiatives.

Although the application of machine learning in ecological prediction has been increasingly explored, studies specifically addressing sea turtle hatching success remain scarce. Prior research employing Multiple Linear Regression (MLR) and Decision Tree (DT) algorithms achieved Root Mean Square Error (RMSE) values of 3.96 (training) and 4.95 (testing) for MLR, and 4.29 (training) and 4.82 (testing) for DT. Despite acceptable accuracy, these models failed to incorporate temporal intervals between nesting and hatching events [6]. Conversely, conventional LSTM models have achieved up to 97.13% accuracy in smart grid forecasting tasks [15]. Demonstrating their superiority in handling large-scale, long-term time-series data for predictive planning in energy systems [16]. Empirical comparisons also show that LSTM consistently outperforms ARIMA models in capturing nonlinear, long-term, and seasonal dependencies [17]. Although Transformer-based architectures have recently gained attention, they generally require larger datasets and greater computational resources, making them less suitable for small scale ecological studies. Within the context of sea turtle conservation, there remains a research gap in integrating deep learning based approaches, particularly LSTM, with ecological variables such as sand temperature, humidity, incubation duration, and nesting seasonality. Addressing this gap is crucial for developing data driven predictive models that can simultaneously account for environmental factors and temporal patterns[17] [18]. This study extends existing literature by incorporating key ecological time-series variables including nesting activity, incubation period, environmental conditions, and seasonal nesting trends to accurately predict hatching success rates even under limited data conditions. Unlike prior approaches, the proposed optimized LSTM model is specifically tailored for small-scale, seasonally dependent ecological datasets. Beyond numerical prediction, this model supports evidence-based conservation management and strategic

https://doi.org/10.31849/digitalzone.v16i2. 28856

ecotourism planning. The novelty of this research is applying an LSTM approach models within the context of sea turtle conservation in the Anambas Islands, an area that has received limited scientific attention. This study not only emphasizes predictive accuracy but also demonstrates the practical applicability of LSTM for promoting sustainable marine resource management and enhancing conservation-oriented ecotourism planning.

2. Research Method

The research was conducted using a quantitative research approach [19]. The data were processed through a secondary analysis approach and deep learning modelling, specifically using the LSTM model to predict sea turtle egg hatching. The object of the study was sea turtles, with the variables utilised in the research including nest code, egg-laying date, nest depth, number of eggs, turtle species, tidal distance, hatching date, temperature, and humidity. The research stages included data loading and comprehension, Exploratory Data Analysis (EDA) to observe data structure, patterns, and characteristics, and data splitting for testing purposes, which consisted of training and testing datasets with configurations of 90:10, 80:20, 70:30, and 60:40. The modelling phase employed the LSTM method with time steps of 2, 5, 7, 30, and 45. The model was evaluated using the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) metrics to assess how well the model predicted the outcomes [20], it can be seen in the flow diagram Figure 1.

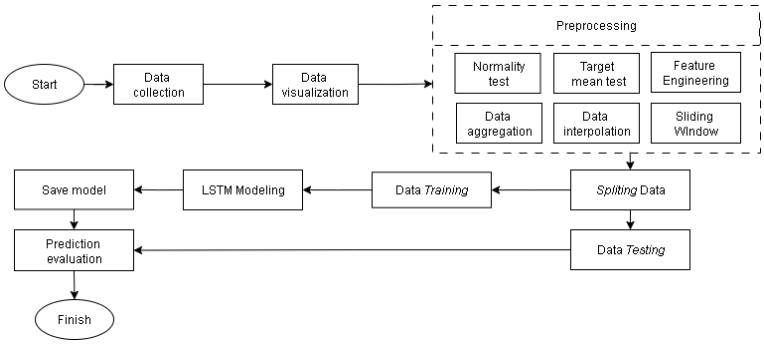

Figure 1: research workflow

Figure 1 illustrates the research workflow for predicting sea turtle hatching success using an LSTM-based framework. The process begins with defining research objectives and collecting both environmental and hatchery data, including temperature, humidity, nesting dates, incubation periods, and hatching success rates. The preprocessing stage ensures data quality through normality testing to Ensures that the data distribution meets the assumptions required for statistical analysis, visualization, aggregation to Integrates multiple data sources into a unified and standardized format, and interpolation to Fills in missing values and smooths irregularities to maintain data continuity and completeness, aiming to produce clean and reliable data for prediction. Feature engineering techniques process transforms raw variables into meaningful features that enhance model performance. Derived features may include time-lagged values, temperature averages, and environmental interaction indices, and the sliding window method are applied to transform the data into supervised sequences suitable for LSTM temporal modeling. The dataset is then divided into training and testing subsets, where the training data are used to train the model and the testing data are used to evaluate its performance. The LSTM model is trained to capture short and long-term temporal dependencies within the data. Once optimal performance is achieved, the model is saved and evaluated using MAE, RMSE, and R^2 metrics to validate its accuracy and generalization capability.

Choosing an appropriate model for ecological time-series prediction is essential to capture both short- and long-term dependencies. While traditional models like ARIMA perform well on linear stationary data, they struggle with the nonlinear and dynamic characteristics of ecological systems [10]. Deep learning models, particularly LSTM networks, have shown

superior performance in forecasting complex time-series data, including environmental and marine ecosystem applications [21] [17]. Compared to GRU and Bi-LSTM, LSTM offers a balanced trade-off between accuracy and computational efficiency. Although Transformer models provide strong predictive capability, their high computational demand limits their use in small-scale ecological studies. Therefore, LSTM is chosen for its robustness in modeling temporal dependencies from limited ecological datasets while maintaining computational efficiency [18][22].

2.1 Dataset Description

Data were systematically collected through the identification and selection of credible sources. The primary dataset was obtained from the monitoring records of LKKPN Pekanbaru, while environmental variables, including temperature and humidity, were retrieved from the official BMKG open-access portal https://dataonline.bmkg.go.id/data-harian. The environmental records were selected for their relevance to sea turtle nesting and hatching success. Each source was critically evaluated to ensure data quality and research suitability in terms of completeness, temporal coverage, and consistency. Only datasets that met these standards were retained for analysis, as summarized in **Table 1.**

Table 1. The initial research data were compiled from multiple credible sources

Kode sarang[1]	Tgl bertelur[1]	Kedalaman sarang[1]	Jumlah telur[1]	Jenis Penyu[1]	Pasang surut[1]	Tgl menetas[1]	Suhu[2]	Kelembaban[2]
R1	4-Mar-22	30	150	sisik	15	25-Apr-22	26.2	88
R1s	12-Mar-22	30	129	sisik	15	24-Apr-22	26.7	82
R2	14-Mar-22	41	90	hijau	15	28-Apr-22	26.8	81
...	...	...	...	...	...	...	...	...
r39	31-May-24	37	140	sisik	22	18-Jul-24	30	80

1044 rows × 9 columns

[1]Source: Sea turtle monitoring data provided by LKKPN Pekanbaru
[2]Source: Environmental data retrieved from the BMKG open-access online portal

Each data source was critically evaluated to ensure quality, completeness, and consistency. Only qualified datasets were included in the analysis. After selection, all data were harmonized into a unified structured format to support preprocessing, exploratory analysis, and modelling. This integrated dataset served as the empirical basis for predictive modelling, comprising 1,044 records with nine variables: nest code, laying date, nest depth, egg count, species, tidal range, hatching date, temperature, and humidity.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to provide an overall visual overview of the dataset, as well as to examine and interpret the relationships between its variables [23]. Exploratory Data Analysis was selected due to its ability to present data in a visually appealing and easily interpretable form for the audience[24]. A comprehensive visualization of the dataset is provided in the following **Figure 2**.

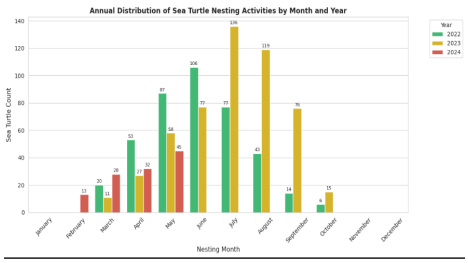

Figure 2a: Distribution of Turtle nesting the Research Dataset

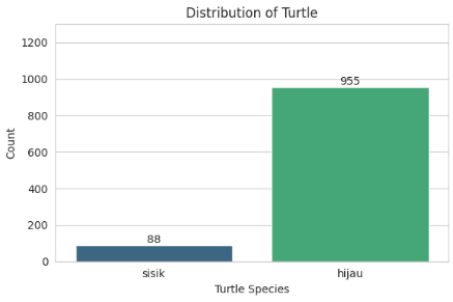

Figure 2b: Distribution of Turtle Species in the Research Dataset

	kedalaman_sarang	jumlah_telur	pasut	suhu	kelembaban
count	1044.000000	1044.000000	1044.000000	1044.000000	1044.000000
mean	61.531609	92.732759	17.526149	28.625096	80.791188
std	19.040200	23.709702	3.253153	1.061152	4.069445
min	17.000000	21.000000	6.000000	25.600000	73.000000
25%	47.750000	78.000000	15.000000	28.000000	78.000000
50%	60.000000	91.000000	18.000000	28.700000	80.000000
75%	70.000000	105.000000	20.000000	29.400000	84.000000
max	125.000000	187.000000	31.000000	30.900000	94.000000

Figure 2c: Distribution of Turtle Species in the Research Dataset

Figure 2a This figure illustrates annual variations in sea turtle nesting activity from 2022 to 2024, showing distinct temporal patterns. Nesting typically increased in March, peaked between May and July, and declined by October, with minimal activity in November and December. The peak occurred in June 2022 (106 nests), July 2023 (136 nests), and May 2024 (45 nests), indicating fluctuations in nesting intensity and seasonality across years

Figure 2b illustrates that the distribution of turtle species is predominantly composed of green turtles. This is attributable to their wider presence within the Anambas Islands and National Marine Protected Area. In terms of physical characteristics, green turtles are generally larger than hawksbill turtles [14]. Subsequently, the distribution of variables across different turtle species can be observed in **Figure 3**.

Figure 2c presents descriptive data analysis using the *describe ()* function, providing key statistical insights such as mean, standard deviation, range, and distribution. This step is essential for LSTM modeling, as it ensures proper normalization or standardization to stabilize gradients, accelerate convergence, and enhance predictive accuracy. Furthermore, this analysis aids in identifying the most influential variables and detecting potential outliers that may degrade model performance. The distribution of data variables across turtle species can be observed in **Figure 4**.

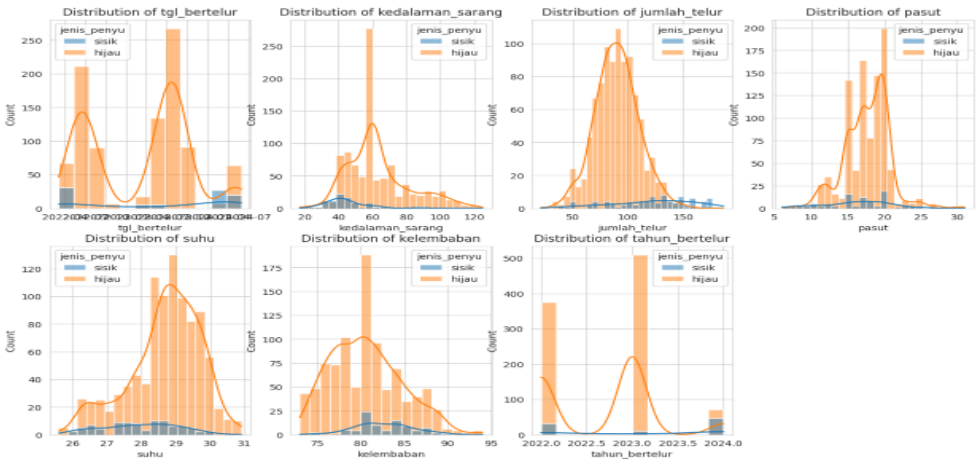

Figure 3: Variable-wise distribution across turtle species

Variable-wise visualizations across turtle species were constructed to assess inter-species variation **Figure 3**. Most variables exhibited similar distributions between green and hawksbill turtles, although several data points fell outside the expected ranges, suggesting the existence of outliers. Detecting these deviations was essential to determine whether the variations stemmed from ecological differences or measurement inconsistencies. Outlier detection was conducted using the Interquartile Range (IQR) method [25] , and the results were visualized through boxplots **Figure 4**.

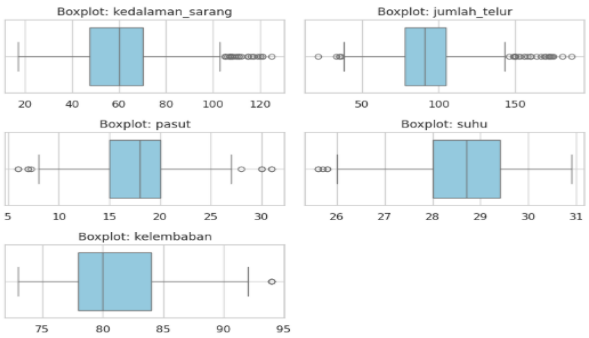

Figure 4: Outlier detection plots for each variable

Outlier detection plots for each variable the plots revealed that nest depth, egg count, and tide variables contained several outliers positioned beyond the whiskers, indicating deviations from normal ranges. These outliers may represent either natural ecological variation or potential recording errors. A detailed summary of outlier thresholds and frequencies by turtle species is presented in **Table 2**, providing a clearer understanding of species-specific variability**.**

Table 2. Outlier summary variable across turtle species

Variables	Turtle Species	Lower Bound	Upper Bound	Count Below Lower Bound	Count Above Upper Bound
Kedalaman sarang	Hijau	18.50	102.50	1	40
	Sisik	20.00	68.00	0	3
Jumlah telur	Hijau	40.75	138.75	7	12
	Sisik	23.75	225.75	0	0
Suhu	Hijau	26.15	31.35	8	0
	Sisik	25.20	30.80	0	0
Kelembaban	Hijau	70.50	90.50	0	10

A further temporal visualization **Figure 5** was developed to explore the recurrence of seasonal nesting behaviors. The results confirmed consistent nesting peaks between May and September, reinforcing the cyclical nature of sea turtle reproduction. However, the 2024 dataset only covered observations up to June, thereby limiting temporal continuity and reducing the completeness of the time series. Recognizing this limitation is vital to ensure that time-dependent features are accurately modeled in the LSTM framework.

The Exploratory Data Analysis stage provides a comprehensive understanding of temporal behaviors, species composition, statistical characteristics, and potential irregularities within the dataset. This analytical stage is conducted to ensure the validity of subsequent modeling processes. The visual and statistical insights obtained through EDA strengthen the methodological foundation of this study by ensuring that the data used are clean, well-structured, and scientifically robust for predicting the hatching success rate of sea turtles using the LSTM model.

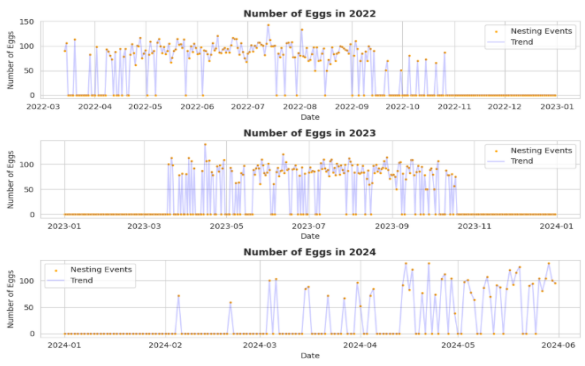

Figure 5: seasonal nesting patterns of sea turtles

2.3 Data Pre-Processing

Data preprocessing was conducted to ensure consistency and quality before model development LSTM. Feature engineering involves the creation, modification, or selection of new variables that more effectively represent the underlying phenomena under investigation. This stage plays a crucial role in enhancing the model's ability to capture complex relationships among variables. Using the following pseudocode df['tgl_menetas'] = pd.to_datetime (df['tgl_menetas']), df['tgl_bertelur'] = pd.to_datetime(df['tgl_bertelur']), and df['lama_inkubasi'] = (df['tgl_menetas'] - df['tgl_bertelur']).dt. days the incubation period was derived by calculating the difference between the hatching date and the laying date.

A normality test was performed to assess whether the data conformed to a normal distribution. This procedure aimed to verify the distributional characteristics of the dataset prior to further statistical analysis. The Shapiro Wilk test was applied for this purpose. The results revealed that the target variable, incubation period, did not exhibit a normal distribution for either turtle species. Specifically, the p-value for hawksbill turtles was 1.6944145415499533e-21, and for green turtles, 8.979710046018405e-12. Both values were considerably lower than the significance threshold ($\alpha = 0.05$), thus leading to the rejection of the null hypothesis (H_0) of normality [26][27] . The spread of the incubation period data for each turtle species is illustrated in **Figure 6**.

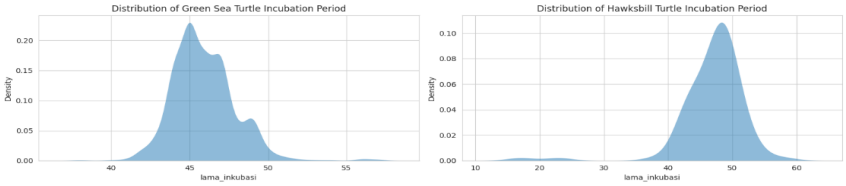

Figure 6: Distribution plots by turtle species

As the incubation period data were non-normally distributed, the Mann–Whitney U test was used to compare hawksbill and green turtles, yielding a p-value of 0.0000007459 ($\alpha = 0.05$), indicating a significant difference [28]. Consequently, the species could not be combined, and hawksbill turtle data were excluded to allow focused aggregation on green turtles, facilitating the construction of a complete time series suitable for LSTM modeling.

To prepare the dataset for LSTM modeling, it was transformed into a time series format using the nesting date as the index. Data for green turtles were aggregated by date to create unique daily records, and missing dates were inserted with zeros to indicate non-nesting days. A binary feature, "nesting status," was added (1 for nesting days, 0 for non-nesting days) to enable the model to learn both nesting and non-nesting patterns, ensuring a complete and chronologically ordered time series illustrated in **Figure 7**.

```
<class 'pandas.core.frame.DataFrame'>
Index: 955 entries, 2 to 1042
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   tgl_bertelur     955 non-null    datetime64[ns]
 1   kedalaman_sarang 955 non-null    int64
 2   jumlah_telur     955 non-null    int64
 3   jenis_penyu      955 non-null    object
 4   pasut            955 non-null    float64
 5   tgl_menetas      955 non-null    datetime64[ns]
 6   suhu             955 non-null    float64
 7   kelembaban       955 non-null    int64
 8   lama_inkubasi    955 non-null    int64
dtypes: datetime64[ns](2), float64(2), int64(4), object(1)
memory usage: 74.6+ KB
```

Data Composition Before Aggregation and Index Transformation

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 810 entries, 2022-03-14 to 2024-05-31
Freq: D
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   kedalaman_sarang 810 non-null    int64
 1   jumlah_telur     810 non-null    int64
 2   pasut            810 non-null    float64
 3   suhu             810 non-null    float64
 4   kelembaban       810 non-null    int64
 5   lama_inkubasi    810 non-null    int64
 6   status_bertelur  810 non-null    int64
dtypes: float64(2), int64(5)
memory usage: 50.6 KB
```

Data Composition After Aggregation and Index Assignment

Figure 7: Data Composition Before and After Aggregation and Index Assignment

Data interpolation was applied to address missing values and maintain the continuity of the time series, ensuring no loss of essential information for model training. A complete date index was generated, and the sliding window method was used to create input sequences capturing temporal dependencies. Time steps of 2, 5, 7, 30, and 45 were tested to represent short-, medium,

and long-term nesting patterns, enabling the LSTM model to effectively learn the temporal dynamics of sea turtle nesting activity for hatching success prediction.

2.4 Spliting Data

Sea turtle hatching prediction was performed using an LSTM model, with the dataset divided into training and testing subsets to evaluate generalization and predictive accuracy. Four data split ratios were tested 60:40, 70:30, 80:20, and 90:10. Each influencing model performance differently. The 60:40 split enabled broader evaluation, 70:30 provided a balanced and reliable configuration, 80:20 enhanced learning depth with limited validation data, and 90:10 maximized training capacity while reducing generalization assessment strength.

2.5 Modeling LSTM

The LSTM model consists of two layers and is trained using time steps of 2, 5, 7, 30, and 45 to capture the natural nesting patterns of sea turtles. Shorter time steps (2 - 7 days) reflect daily and weekly variations, while longer steps (30 - 45 days) capture monthly cycles and full incubation periods. LSTM was chosen for its effectiveness in modeling temporal dependencies and its suitability for small-scale datasets [17].

The optimized model employs 256 units in the first LSTM layer, 32 units in the second, a 0.3 dropout rate, and a single dense output layer. It is trained using the Adam optimizer with a learning rate of 0.0005, a batch size of 32, and up to 100 epochs, with early stopping to mitigate overfitting. The Huber loss function ($\delta = 1.0$) is applied, and performance is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The selection of the LSTM model was conducted using a trial-and-error approach, where each parameter configuration was tested iteratively to identify the optimal model structure. The process was implemented through the following function

```
def train_model (data, split, time_step, epochs=100, batch_size=32):
    X_train = data[split][time_step] ['X_train']
    y_train = data[split][time_step] ['y_train']
    n_features = X_train.shape[2]
    model = create_lstm_model(time_step=time_step, n_features=n_features)
    early_stop = EarlyStopping(
        monitor='val_loss',
        patience=5,
        restore_best_weights=True)
    hist = model.fit(
        X_train, y_train,
        epochs=epochs,
        batch_size=batch_size,
        validation_split=0.2,
        callbacks=[early_stop],
        verbose=0 )
return model, hist
```

The main limitation of the model during the training process lies in the limited availability of training data, which poses challenges in determining the optimal parameters for the LSTM model. However, a practical implication of this research is that conservation area managers can gain valuable insights specifically, that sea turtles are likely to return for nesting approximately seven days after their initial nesting event.

2.6 Model Evaluation

In this study, the model's performance in predicting sea turtle hatching success is evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Both RMSE and MAE serve as indicators of model accuracy, providing insight into the magnitude of prediction errors relative to the actual values [29]. The equations for RMSE and MAE are presented.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - y`i)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi - y`i|$$

Beyond the two evaluation metrics mentioned above, the R^2 score (coefficient of determination) is employed to evaluate the deviation between the predicted and actual values[18]. The equation for R^2 is given.

$$R^2 = \frac{\sum_i (p(xi) - xi)2}{\sum_i (\bar{x} - x_i)2}$$

3. Results and Discussion

3.1 Results

he experimental results, obtained from 810 data points, were evaluated using various data split ratios of 60:40, 70:30, 80:20, and 90:10, as well as time step configurations of 2, 5, 7, 30, and 45. The detailed outcomes of these evaluations are presented in **Table 3.**

Table 3. Configuration data Based on Time Steps and split

Time Steps	Split data	Shape x	Shape y	Time Steps	Split data	Shape x	Shape y
2	60:40	484, 2, 6	484	7	80:20	641, 7, 6	641
	70:30	565, 2, 6	565		90:10	722, 7, 6	722
	80:20	646, 2, 6	646	30	60:40	456, 30, 6	456
	90:10	727, 2, 6	727		70:30	537, 30, 6	537
5	60:40	481, 5, 6	481		80:20	618, 30, 6	618
	70:30	562, 5, 6	562		90:10	699, 30, 6	699
	80:20	643, 5, 6	643	45	60:40	441, 45, 6	441
	90:10	724, 5, 6	724		70:30	522, 45, 6	522
7	60:40	479, 7, 6	479		80:20	603, 45, 6	603
	70:30	560, 7, 6	560		90:10	684, 45, 6	684

Table 3 illustrates that the data distribution varies with each time step configuration. The size of the training set depends on the chosen time step, as the model utilizes this window to predict the next data point. A higher training percentage yields more data for model learning and less for testing. The combination of time steps and data split ratios produces 20 distinct LSTM models, each representing a unique configuration. The detailed predictive performance of these models is summarized in **Table 4**.

Table 4. Model performance in predicting based on data splitting and time step configurations

Model Name	Split	Time Step	RMSE	MAE	R^2
model9	60:40	7	17.89947	8.672685	0.339702
model5	60:40	5	20.06034	13.32657	0.173601
model10	70:30	7	17.03494	9.626145	0.162382
model1	60:40	2	20.38972	11.32094	0.150654
...	...	...	...	...	...
model20	90:10	45	38.36831	32.69637	-2.54184

Table 4 presents the evaluation of 20 LSTM models with varying time step lengths and data split ratios, assessed using RMSE, MAE, and R^2. Each model was designed to capture the underlying temporal patterns of sea turtle egg hatching data. Overall, the four most accurate models in capturing the incubation patterns were those trained on larger proportions of data (i.e., 60:40 and 70:30 training-to-testing ratios). This is supported by their lower RMSE and MAE values and relatively higher R^2 scores, indicating better predictive reliability. The models are ranked in descending order based on their R^2 values, reflecting the extent to which each model explains the variance in the target variable.

Model9, utilizing a 60:40 data split and a time step of 7, achieved the highest performance with an MAE of 8.6727, RMSE of 17.8995, and R^2 of 0.3397. Model5, with the same data ratio and a time step of 5, followed with an MAE of 13.3266, RMSE of 20.0603, and R^2 of 0.1736. Model10, built with a 70:30 split and a time step of 7, resulted in an MAE of 9.6261, RMSE of 17.0349, and R^2 of 0.1624. Model1, employing a 60:40 split with a time step of 2, produced an MAE of 11.3209, RMSE of 20.3897, and R^2 of 0.1507.

In contrast, the remaining models demonstrated suboptimal performance, particularly those configured with smaller training sizes (e.g., 90:10 splits). These models exhibited R^2 values near zero or negative, suggesting poor generalization capabilities. Notably, Model20, which employed a 90:10 split and a time step of 45, recorded the lowest performance with an R^2 of -2.5418. Such findings indicate that insufficient training data severely limits the model's ability to learn meaningful temporal dependencies and to capture the underlying variability in incubation duration, often performing worse than a simple mean-based baseline. The prediction results of the model can be seen in **Figure 9**.

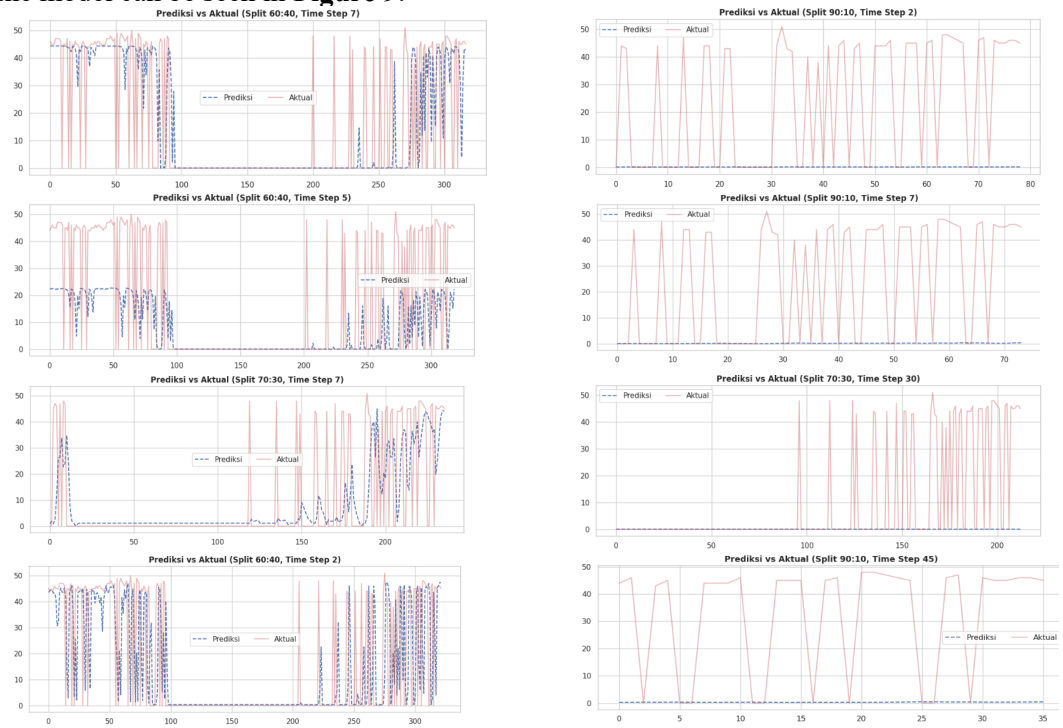

Figure 9: Visualization of the prediction results and actual data on the top mode

Figure 9 illustrates the comparison between actual observations and LSTM model predictions, demonstrating the influence of short and long-time step configurations. The model shows competence in capturing seasonal hatching patterns, though performance declines in sequences with extended temporal dependencies. The predicted values of the incubation period, as produced by the model, are summarized in **Table 5**

Table 5. Comparative analysis of actual versus predicted outcomes from top LSTM models under each time step and data partition setting

Time Steps 7, 60:40		**Time Steps 5, 60:40**		**Time Steps 7, 70:30**		**Time Steps 2, 60:40**	
y_test	**y_pred**	**y_test**	**y_pred**	**y_test**	**y_pred**	**y_test**	**y_pred**
46	44.330379	44	22.36551	0	0.932037	45	43.526409
45	44.374462	45	22.35416	0	1.819896	44	44.229607
45	44.38007	46	22.40974	45	0.879532	45	44.346283
45	44.356724	45	22.43518	47	2.461845	44	44.276794

Time Steps 7, 60:40		**Time Steps 5, 60:40**		**Time Steps 7, 70:30**		**Time Steps 2, 60:40**	
y_test	**y_pred**	**y_test**	**y_pred**	**y_test**	**y_pred**	**y_test**	**y_pred**
...	...	...	...	...	...	...	...
45	42.924183	45	22.35974	45	44.103996	45	47.340778

Based on **Table 4**, the comparison between the actual (*y_test*) and predicted (*y_pred*) values across various LSTM configurations time steps of 2, 5, 7, and 45 with different data splits demonstrate the model's ability to generalize temporal patterns, though some bias remains. The configuration with 7-time steps and a 60:40 split achieved the highest predictive accuracy, with minor deviations (≈3.93) between actual and predicted values, indicating effective temporal pattern recognition. Conversely, the 5 time-step configurations with the same split produced structured yet inaccurate predictions, suggesting the model's inability to capture temporal dependencies despite sufficient training data. The 7 time-step and 70:30 split configuration failed to generalize effectively, producing unrealistically low predictions due to limited exposure to data variability. Interestingly, the 2 time-step and 60:40 split configuration yielded predictions closely aligned with actual values, implying that short time windows can effectively capture short-term incubation dynamics. However, its low R^2 (0.15) and high MAE (11.32) indicate that while predictions are stable, the model lacks precision in representing the full variance of incubation duration.

3.2 Discussion

Handling sea turtle hatching datasets with seasonal time series patterns poses a major challenge in LSTM modeling. Interpolation techniques are applied to preserve data continuity without introducing artificial values, reflecting periods without nesting or hatching activity. To enhance contextual learning, a "nesting status" variable was introduced, where 0 indicates no hatching and 1 indicates hatching occurrence. The LSTM model effectively captures temporal patterns during training; however, limited test data constrain its predictive performance on unseen samples. As shown in **Figure 9**, where the model is still able to interpret and process an input value of 0 indicating a day without hatching activity and subsequently responds appropriately when it identifies the presence of such activity. The results of this study reveal that the performance of the LSTM model in predicting the incubation period of sea turtle eggs is highly dependent on two key factors: the data split configuration used for training and prediction, and the temporal pattern of nesting behavior, which directly influences the hatching period. The model interprets these behavioral patterns as time steps.

The model demonstrates a better capability in recognizing temporal patterns when configured with longer time steps, provided that there is sufficient training data to support such a configuration. In this research, a time step length of 7 (corresponding to a weekly pattern) combined with a 60:40 training-to-testing data ratio was found to yield the best generalization performance for seasonal temporal trends. These findings underscore the critical role of test data availability in enabling the model to generalize effectively from the training data. This highlights a broader implication: model reliability in time series forecasting is not solely dependent on architecture, but also on the representativeness and completeness of the dataset used for both training and evaluation.

The optimal configuration identified in this study involves a time step of 7 reflecting the weekly hatching pattern of sea turtle eggs and a data split ratio of 60:40. This configuration produces consistent, stable, and accurate predictions, with a moderate error range. In practical and operational terms, this means that the difference between actual and predicted incubation periods generally falls within an acceptable deviation of ± 3 to a maximum of ± 5 days. As such, the model's predictions demonstrate a tolerable level of deviation from real values. Under this configuration, the model achieved a Mean Absolute Error (MAE) of 8.67, a Root Mean Squared Error (RMSE) of 17.89, and a coefficient of determination (R^2) of 0.34. These evaluation metrics suggest that the model is capable of identifying, learning, and generalizing the main temporal patterns present in the incubation dataset, even though some minor fluctuations remain

imperfectly captured.The findings emphasize that selecting appropriate time steps in LSTM modeling significantly influences the model's ability to understand the structure of temporal data particularly in the context of seasonal and medium-term forecasting [17].

The configuration using a time step of 5 with a 60:40 data split reflects a condition in which the model tends to undervalue its predictions. This indicates that a shorter time step fails to adequately capture the underlying patterns and the complexity of temporal relationships within the data. This limitation is evident from the evaluation metrics, which show a Mean Absolute Error (MAE) of 13.33, a Root Mean Squared Error (RMSE) of 20.06, and a coefficient of determination (R^2) of 0.17. Such performance suggests that the LSTM model, which is highly sensitive to the sliding window size, requires sufficient temporal input in order to effectively learn and generalize seasonal patterns. Inadequate input length restricts the model's ability to grasp recurring trends, resulting in less accurate predictions [30].

The configuration with a time step of 7 and a 70:30 data split resulted in predictions that were not representative of the actual values. Although the model used a relatively long time step, the limited amount of training data hindered its ability to learn the underlying data structure effectively, leading to underfitting. This is reflected in the evaluation metrics: a Mean Absolute Error (MAE) of 9.63, a Root Mean Squared Error (RMSE) of 17.03, and a coefficient of determination (R^2) of 0.16. Meanwhile, the configuration with a time step of 2 and a 60:40 split produced predictions that were consistent with the overall trend of the actual data, but with low accuracy, as indicated by a low R^2 value of 0.15. This suggests that the model was able to predict values close to the average but failed to capture the full variability of the data. These findings are consistent with previous research, which highlights the importance of balancing time step configuration with adequate training data volume to ensure that LSTM models can effectively learn temporal structures and produce reliable predictions [31] This supports the notion that the number of short-term memory units in an LSTM is not, by itself, a sufficient indicator of the model's ability to capture fluctuations or complex dynamics within sequential data, unless it is supported by an adequately wide time step window..

The residual analysis, presented in **Figure 10**, shows a weak correlation between predicted incubation periods and prediction errors. Most predictions cluster between 30 and 45 days, closely matching actual values, indicating good model performance within the dataset's central range. Larger residuals occur in rare cases where predictions fall below 10 days or exceed 45 days, reflecting the model's limited generalization beyond typical patterns. Statistical evaluation further reveals higher accuracy around the second quartile (Q2), where error margins are smaller, demonstrating the model's strength in capturing average temporal dynamics of incubation periods.

However, the model struggles to predict outlier cases with high precision. This limitation is consistent with the nature of LSTM models, which depend on patterns present in the training data. When such extreme patterns are underrepresented, the model fails to form reliable generalizations. These extreme residuals may also be explained by environmental variability, particularly changes in temperature and humidity. For instance, higher temperatures during dry seasons may accelerate embryonic development, leading to earlier-than-average hatching, while during rainy seasons, elevated humidity may slow down the process, resulting in delayed hatching. Additionally, nest microclimates such as those located in shaded or protected areas can further influence the rate of heat exchange, affecting incubation duration. Understanding this residual pattern is essential for interpreting the limitations of the model and for proposing future improvements, such as incorporating environmental features (e.g., soil temperature, rainfall, nest exposure) into the input variables to enhance model performance on extreme cases [32] The residual analysis is illustrated in **Figure 10**.

https://doi.org/10.31849/digitalzone.v16i2. 28856

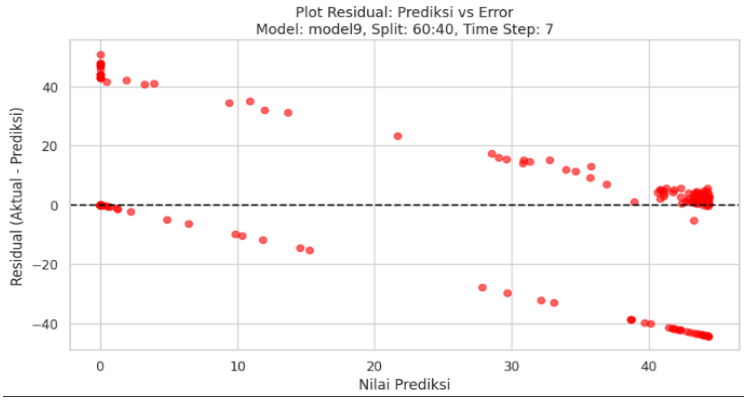

Figure 10: Residual plot

The main limitation of this study is the limited training and testing data, which restricts the model's ability to learn temporal patterns effectively. Future research should address this by expanding the dataset and employing more comprehensive monitoring data to improve model generalization and predictive performance using advanced deep learning methods.

4. Conclusions

The LSTM model with a 7-day time step and a 60:40 training-testing ratio achieved the highest predictive accuracy, reflecting the natural 7 - 14-day nesting cycle of sea turtles. The model effectively captured temporal incubation patterns, with residuals centered near zero, indicating reliable prediction performance. Although minor deviations occurred at extreme values, most predictions fell within acceptable error margins. These results highlight the model's strong generalization capability and its potential to support data-driven conservation planning, particularly for optimizing nest monitoring, predator management, and sustainable ecotourism development in coastal ecosystems.

The optimized LSTM model can serve as a predictive tool for sea turtle conservation monitoring, particularly within the Anambas Islands Marine Protected Area. To enhance predictive accuracy, future developments should incorporate internal environmental factors at nesting sites such as sand temperature, humidity, and predator activity along with the expansion of research data, as the current model relies solely on external environmental variables as input. Integrating this model into web-based or mobile spatial information systems could improve the efficiency of turtle monitoring data collection and support decision-making processes for conservation managers and ecotourism stakeholders. Further research is recommended to explore hybrid modeling approaches to enhance both the accuracy and interpretability of the predictive framework.

References

[1] L. P. LKKPN Pekanbaru, "Monitoring Penyu di Kawasan Kosnervasi Kepulauan Anambas Tahun 2024, [Internal report]. Available: LKKPN Pekanbaru.

[2] T. Z. A. E. Hamino, I. N. Y. Parawangsa, L. A. Sari, and S. Arsad, "Efektifitas Pengelolaan Konservasi Penyu di Education Center Serangan , Denpasar Bali," *J. Mar. Coast. Sci. Vol.*, vol. 10, no. 1, pp. 18–34, 2021, [Online]. Available: https://e-journal.unair.ac.id/JMCS/article/download/25604/13512

[3] R. R. K. Sinaga, A. Hanif, F. Kurniawan, S. Roni, D. Y. W. Laia, and J. R. Hidayati, "Tingkat Keberhasilan Penetasan Telur Penyu Hijau (Chelonia mydas) dan Penyu Sisik (Eretmochelys imbricata) Di Pulau Mangkai Kepulauan Anambas," *J. Mar. Res.*, vol. 13, no. 1, pp. 92–99, 2024,https://doi.rg/10.14710/jmr.v13i1.38531.

[4] A. Hanif, H. Damanhuri, S. Suparno, and M. U. Rusli, "Tingkat Penetasan Penyu Hijau di

Pulau Pandan Kawasan Konservasi Pulau Pieh, Sumatera Barat," *J. Akuatiklestari*, vol. 6, no. 1, pp. 1–9, 2022, https://doi.org/10.31629/akuatiklestari.v6i1.4696

[5] Ikha Safitri, "Monitoring Penyu sebagai Upaya dalam Pengelolaan KKP3K Paloh Kalimantan Barat," *JurnalPengabdian Kpd. Masy. Nusant.*, vol. 5, no. 1, p. 120, 2024.

[6] A. Agustriono, S. Rapindra, and R. Rahmaddeni, "Komparasi Multiple Linear Regression dan Decision Tree dalam Memprediksi Penetasan Penyu Jenis Chelonioidea Sp di Pulau Mangkai," vol. 14, no. 1, pp. 9–17, 2024. https://ejurnal.umri.ac.id/index.php/JIK/article/view/6844

[7] L. P. LKKPN Pekanbaru, "Laporan Monitoring Penyu Kawasan Konservasi Pieh dan Kawasan Konservasi Anambas Tahun 2022," 2022. [Internal report]. Available LKKPN Pekanbaru.

[8] L. P. LKKPN Pekanbaru, *Laporan Monitoring Penyu 2024*. 2024. [Internal report]. Available: LKKPN Pekanbaru.

[9] L. P. LKKPN Pekanbaru, "Laporan Monitoring Penyu Kawasan Konservasi Anambas Tahun 2023," 2023. [Internal report]. Available: LKKPN Pekanbaru.

[10] R. C. Edwards, B. J. Godley, and A. Nuno, "Exploring connections among the multiple outputs and outcomes emerging from 25 years of sea turtle conservation in Northern Cyprus," *J. Nat. Conserv.*, vol. 55, no. December 2019, p. 125816, 2020, htts://doi.org/10.1016/j.jnc.2020.125816.

[11] L. P. LKKPN Pekanbaru, *Laporan Kinerja Tahunan 2024 LKKPN Pekanbaru*, vol. 11, no. 1. 2024.

[12] A. Khumaidi, R. Raafi'udin, and I. P. Solihin, "Pengujian Algoritma Long Short Term Memory untuk Prediksi Kualitas Udara dan Suhu Kota Bandung," *J. Telemat.*, vol. 15, no. 1, pp. 13–18, 2020, https://doi.org/10.61769/telematika.v15i1.340

[13] N. Yudistrira *et al.*, *Prediksi deret waktu menggunakan Deep Learning*, I. Indonesia: UB Pres, 2023. [Online]. Available: https://ubpress.ub.ac.id/?p=4433

[14] Departemen Kelautan dan Perikanan, Pedoman Teknis Pengelolaan Konservasi Penyu. 2009, 2009. https://perpustakaan.kkp.go.id/knowledgerepository/index.php?p=show_detail&id=13252

[15] M. Alazab, S. Khan, S. S. R. Krishnan, Q. V. Pham, M. P. K. Reddy, and T. R. Gadekallu, "A Multidirectional LSTM Model for Predicting the Stability of a Smart Grid," *IEEE Access*, vol. 8, pp. 85454–85463, 2020, https://doi.org/10.1109/ACCESS.2020.2991067.

[16] A. Muneer, R. F. Ali, A. Almaghthawi, S. M. Taib, A. Alghamdi, and E. A. A. Ghaleb, "Short term residential load forecasting using long short-term memory recurrent neural network," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 5, pp. 5589–5599, 2022, http://doi.org/10.11591/ijece.v12i5.pp5589-5599

[17] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," *Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018*, pp. 1394–1401, 2018, doi: 10.1109/ICMLA.2018.00227.

[18] J. Wen, J. Yang, B. Jiang, H. Song, and H. Wang, "Big Data Driven Marine Environment Information Forecasting: A Time Series Prediction Network," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 1, pp. 4–18, 2021, https://doi.org/10.1109/TFUZZ.2020.3012393.

[19] Y. Rifa'i, "Analisis Metodologi Pengumpulan Data di Penelitian Ilmiah," *Cendekia Inov. Dan Berbudaya*, vol. 1, no. 1, pp. 31–37, 2023. https://doi.org/10.59996/cendib.v1i1.155

[20] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," no. 2, pp. 5481–5487, 2022. https://doi.org/10.5194/gmd-15-5481-2022

[21] Y. P. Chen *et al.*, "Real-time decision-making for Digital Twin in additive manufacturing with Model Predictive Control using time-series deep neural networks," *J. Manuf. Syst.*, vol. 80, no. March, pp. 412–424, 2025, https://doi.org/10.1016/j.jmsy.2025.03.009.

[22] Z. M. Shaikh and S. Ramadass, "Unveiling deep learning powers: LSTM, BiLSTM, GRU,

BiGRU, RNN comparison," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 35, no. 1, pp. 263–273, 2024, https://doi.org/10.11591/ijeecs.v35.i1.pp263-273.

[23] D. W. A. R. F. E. P. A. Satyanarayan and A, "Charting EDA: Characterizing Interactive Visualization Use in Computational Notebooks with a Mixed-Methods Formalism," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, pp. 1191–1201, 2025, https://doi.org/10.1109/TVCG.2024.3456217.

[24] A. Wibowo, "Analisa Dan Visualisasi Data Penjualan Menggunakan Exploratory Data Analysis Pada PT. Telkominfra," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, pp. 2292–2304, 2022, https://jurnal.mdp.ac.id/index.php/jatisi/article/view/2737

[25] Ceballos, *Scikit-Learn Decision Trees Explained*. 2019. [Online]. Available: https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d

[26] K. Rani Das, "A Brief Review of Tests for Normality," *Am. J. Theor. Appl. Stat.*, vol. 5, no. 1, p. 5, 2016, https://doi.org/10.11648/j.ajtas.20160501.12.

[27] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Ann. Card. Anaesth.*, vol. 22, no. 1, pp. 67–72, 2019, https://doi.org/10.4103/aca.ACA_157_18.

[28] M. Aslam and M. Sattam Aldosari, "Analyzing alloy melting points data using a new Mann-Whitney test under indeterminacy," *J. King Saud Univ. - Sci.*, vol. 32, no. 6, pp. 2831–2834, 2020, doi: 10.1016/j.jksus.2020.07.005.

[29] H. Abbasimehr, M. Shabani, and M. Yousefi, "An optimized model using LSTM network for demand forecasting," *Comput. Ind. Eng.*, vol. 143, no. July 2019, p. 106435, 2020, https://doi.org/ 10.1016/j.cie.2020.106435.

[30] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Expert Syst. Appl.*, vol. 140, 2020, https://doi.org/10.1016/j.eswa.2019.112896.

[31] F. A. Gers and F. Cummins, "A critique of neoclassical macroeconomics," *Choice Rev. Online*, vol. 27, no. 09, pp. 27-5238-27–5238, 1990, https://doi.org/10.5860/choice.27-5238

[32] Sheavtiyan, T. R. Setyawati, and I. Lovadi, "Tingkat Keberhasilan Penetasan Telur Penyu Hijau (Chelonia Mydas, Linnaeus 1758) di Pantai Sebubus, Kabupaten Sambas," *J. Protobiont*, vol. 3, no. 1, pp. 46–54, 2014. https://doi.org/10.26418/protobiont.v3i1.4581