Imputation of missing microclimate data of coffee-pine agroforestry with machine learning

Heru Nurwarsito [a,1,*], Didik Suprayogo [b,2], Setyawan P. Sakti [c,3], Cahyo Prayogo [d,4], Novanto Yudistira [d,5], Muhammad Rifqi Fauzi [d,6], Simon Oakley [e,7], Wayan Firdaus Mahmudy [d,8]

[a] Faculty of Agriculture, University of Brawijaya, Malang, Indonesia
[b] Faculty of Agriculture, University of Brawijaya, Malang, Indonesia
[c] Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia
[d] Faculty of Computer Science, University of Brawijaya, Malang, Indonesia
[e] Lancaster Environment Centre, UK Centre for Ecology & Hydrology, United Kingdom
[1] heru@ub.ac.id; [2] suprayogo@ub.ac.id; [3] sakti@ub.ac.id; [4] c.prayogo@ub.ac.id; [5] yudistira@ub.ac.id; [6] mrifqifauzi@student.ub.ac.id; [7] soak@ceh.ac.uk; [8] wayanfm@ub.ac.id
* corresponding author

ARTICLE INFO

Article history
Received November 5, 2023
Revised December 18, 2023
Accepted December 27, 2023
Available online December 29, 2023

Keywords
Microclimate data
Interpolation
Shifted interpolation
K-nearest neighbors (KNN)
Linear regression

ABSTRACT

This research presents a comprehensive analysis of various imputation methods for addressing missing microclimate data in the context of coffee-pine agroforestry land in UB Forest. Utilizing Big data and Machine learning methods, the research evaluates the effectiveness of imputation missing microclimate data with Interpolation, Shifted Interpolation, K-Nearest Neighbors (KNN), and Linear Regression methods across multiple time frames - 6 hours, daily, weekly, and monthly. The performance of these methods is meticulously assessed using four key evaluation metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The results indicate that Linear Regression consistently outperforms other methods across all time frames, demonstrating the lowest error rates in terms of MAE, MSE, RMSE, and MAPE. This finding underscores the robustness and precision of Linear Regression in handling the variability inherent in microclimate data within agroforestry systems. The research highlights the critical role of accurate data imputation in agroforestry research and points towards the potential of machine learning techniques in advancing environmental data analysis. The insights gained from this research contribute significantly to the field of environmental science, offering a reliable methodological approach for enhancing the accuracy of microclimate models in agroforestry, thereby facilitating informed decision-making for sustainable ecosystem management.

1. Introduction

Micro-climate data plays a crucial role in agroforestry management, particularly in coffee-pine ecosystems. Tropical agroforestry systems, such as those involving coffee and *Faidherbia albida* trees, aim to mitigate extreme temperatures, highlighting the significance of micro-climate data in designing resilient and climate-smart farming systems [1]. Additionally, the role of agroforestry systems as a climate change mitigation strategy has been emphasized, further underlining the importance of micro-climate data in such ecosystems [2]. Furthermore, the agroforestry system of coffee cultivation in pine forests has been recognized for its essential role in providing ecosystem services, including habitat for wildlife, carbon storage, and sequestration [3]. The use of UAVs for vegetation monitoring in agroforestry applications demonstrates the increasing suitability of advanced technologies in managing agroforestry

https://doi.org/10.26555/ijain.v10i1.1439

systems, emphasizing the need for accurate micro-climate data [4]. The study further examines the complexities inherent in the social-ecological relationships, focusing on the intention of Indonesian coffee farmers to adopt a tree canopy trimming technique within pine-based agroforestry systems. This highlights the imperative for precise microclimate data to inform the implementation of such practices [5].

The collection of microclimate data, encompassing temperature, humidity, and light intensity, presents significant challenges, leading to gaps in existing datasets and necessitating the use of imputation techniques. The gaps in climate change policies, particularly concerning water-related aspects, as emphasized in the AR6 WG1 report, underscore the challenges associated with acquiring comprehensive and accurate microclimate data [6]. Moreover, the need for further research to understand the biodiversity of bacteria in the coffee rhizosphere and their resulting effects highlights the current gaps in knowledge regarding the complex ecological factors that affect microclimatic conditions within agroforestry systems [7]. Moreover, the research focused on methods for completing datasets to detect contamination in groundwater reserves underscores the importance of data imputation techniques in filling voids in microclimatic information. This underscores the prevalent nature of challenges associated with missing data in environmental studies [8]. Additionally, the examination of missing data in spatiotemporal datasets underscores the inescapable occurrence of data gaps in environmental monitoring networks. This further accentuates the imperative for robust imputation methods to effectively address these deficiencies in microclimate data [9].

To address the challenge of missing microclimate data in coffee-pine agroforestry in the UB Forest, it is essential to consider robust imputation methods to recover the missing data. This research shows the potential of unmanned aerial vehicles (UAVs) for vegetation monitoring, providing an efficient alternative to manual field work and highlighting the feasibility of using advanced technologies for data collection in agroforestry systems [4]. Furthermore, research highlighting sophisticated imputation methods based on similarity learning underlines the capability of iterative imputation techniques to interpolate missing values, utilizing the overarching correlation structure within chosen records. This capability proves beneficial for the restoration of missing microclimate data [10]. Moreover, the selection of the imputation method based on the characteristics of the dataset offers insights into the application of donor-based or model-based imputation to tackle missing data. This consideration may be particularly pertinent in the context of recovering microclimate data in the UB Forest [11]. Additionally, this research introduces a random pruning and replacement method for known values to compare missing data imputation models. This approach offers a practical means to evaluate and choose appropriate imputation models for incomplete datasets, potentially proving valuable in the context of recovering lost microclimate data in the UB Forest [12].

A review of existing methods for data imputation reveals a diverse range of approaches and their associated limitations. The approach for integrating proteomic datasets with effective management of missing values exhibits enhanced efficacy in identifying significant proteins, surpassing traditional imputation methods in performance [13]. However, this research scrutinized various imputation and regression procedures for correcting missing values in health records. It underscored the importance of striking a balance between regressor performance and computational cost [14]. Anomaly detection frameworks for wearable device data, emphasizing the increasing use of wearable devices in clinical studies, demonstrate the need for robust imputation methods in this domain [15]. Furthermore, proposes a random pruning and replacement method of known values to compare missing data imputation models, providing insight into the behavior of different imputation methods for incomplete air quality time series [12]. These studies collectively underscore the need for efficient, accurate, and domain-specific imputation methods to address missing data across various fields, from proteomics to health records and environmental sensing.

To address the problem of imputation or recovery of missing microclimate data in coffee-pine agroforestry in the UB forest, machine learning techniques such as Interpolation, Shifted Interpolation, KNN, and Linear Regression can be employed. The missing gaps in the acquired data are crucial indicators of data reliability, and reducing these gaps is essential for ensuring good quality data [16]. Various techniques exist for addressing missing values, such as regression, which are selected based on the characteristics of the data and other factors like precision and accuracy [16]. Additionally, it is important to consider the temporal components in the data, as linear methods often neglect these components [17]. Furthermore, the application of machine learning methodologies, including deep learning and ensemble learning, has been suggested for the imputation or estimation of missing data, offering potential advantages in this context [18], [19]. Moreover, the application of machine learning methods, such as support vector machines and random forests, has been successful in spatial interpolation, which can be relevant for addressing missing microclimate data [20]. These approaches can contribute to the development of a robust framework for imputing or recovering missing microclimate data in the coffee-pine agroforestry in the UB forest, ultimately leading to a more comprehensive and complete dataset for analysis and decision-making.

Machine learning models play a crucial role in various fields, including computer science, mathematics, and artificial intelligence. In the context of methodology, several machine learning models are commonly used, each with its unique characteristics and applications. Interpolation, Shift Interpolation, KNN, and Linear Regression are among the prominent models used. Interpolation techniques, such as those used in video frame interpolation, leverage convolutional neural networks to estimate multiple intermediate frames, enabling the preservation of spatial coherence and the synthesis of high-quality video frames [21], [22]. Additionally, the use of adaptive two-dimensional autoregressive modeling and soft-decision estimation has been proposed for image interpolation, resulting in improved spatial coherence and subjective visual quality of interpolated images [23]. Furthermore, the exhibited precision and reliability of the electron-positron equation of state, dependent on table interpolation of the Helmholtz free energy, highlight the efficacy of the bi-quintic Hermite polynomial as an interpolating function [24].

The research employs Interpolation, Shifted Interpolation, the KNN method, and Linear Regression for imputation of missing data in coffee-pine agroforestry systems. Interpolation and Shifted Interpolation create a functional relationship between known data points, with the latter adding a temporal dimension to capture cyclical microclimate data [25]. The KNN method emphasizes similarity in conditions for accurate imputation, particularly for capturing rapid changes in microclimate data [26]. Linear Regression is identified for its robustness and accuracy, handling diverse microclimate datasets over various temporal scales, making it a versatile tool for short-term and long-term data analysis [27]. Linear regression models have also been extensively employed in various fields. They have been used in software estimation, judgment modeling, and short-term load forecasting, showcasing their versatility and applicability in different domains [28]–[30].

Additionally, the employment of linear regression in process-tracing models of judgment has been underscored, accentuating its ability to delineate underlying processes across various degrees of generality [29]. These methods collectively contribute to a nuanced understanding of microclimate dynamics, equipping practitioners with comprehensive and reliable data for informed decisions, thereby enhancing the sustainability and productivity of agroforestry ecosystems. The incorporation of these machine learning methodologies marks a substantial advancement in the utilization of data science for environmental preservation.

The main goals of this research are to investigate and assess the effectiveness of different Data Imputation Methods in the context of restoring lost or missing data. This improved data integrity is essential for developing more reliable ecological models, making informed decisions on sustainable agroforestry practices, and designing effective environmental conservation strategies. The research contributes to the field by providing a methodological framework that can be applied to similar ecological data challenges, ultimately resulting in Imputation promoting resilience and sustainability in agroforestry ecosystems. Through this work, practitioners and researchers gain access to enhanced tools for predictive analysis, enabling proactive management of coffee-pine agroforestry landscapes in the face of climate variability and change. The focus of this research are models such as Interpolation, Shifted Interpolation, K-Nearest Neighbors (KNN), and Linear Regression. Each of these models embodies a distinct strategy for tackling the prevalent issue of missing data in datasets. Through the application of these models, the research seeks to not only bridge the gaps in data but also to evaluate and compare their performance in aspects of accuracy, efficiency, and appropriateness for various types of datasets. This comparative analysis is designed to offer insights into the strengths and weaknesses of each model, thereby assisting practitioners in choosing the most suitable method for their specific data imputation requirements. Additionally, the research seeks to contribute to the broader understanding of machine learning techniques in data preprocessing, an essential step in ensuring the quality and reliability of data-driven insights and decisions.

The significance of this research is anchored in the potential benefits that accurate imputation of micro-climate data can bring to agroforestry management and environmental research. Micro-climate data plays a crucial role in understanding local climatic variations, which are vital for effective agroforestry practices. Accurately imputed data can enhance the prediction of climate-related events, aiding in the development of more resilient agroforestry systems. This is particularly critical in the face of global climate change, where precise local climate information is essential for adapting farming practices, selecting appropriate crop varieties, and managing natural resources sustainably. Furthermore, in the domain of environmental research, reliable micro-climate data is indispensable for studying ecological patterns, assessing environmental changes, and forecasting future climatic conditions. It can also assist in formulating policies and strategies for environmental conservation and sustainable development.

Therefore, the advancements in micro-climate data imputation methods that this research aims to establish will not only contribute to the technical field of data science but also have a profound impact on agroforestry management practices and environmental research, ultimately supporting the global effort towards sustainable environmental stewardship. This research, which provides insights into the most effective Data Imputation Method for climate data reconstruction, holds the promise of offering instrumental tools for making informed decisions in these critical areas.

2. Method

2.1. Research Area Description

The agricultural forestry region, commonly referred to as Coffee-Pine, within the University of Brawijaya (UB) Forest area, is strategically positioned along the inclines of Mount Arjuno. It can be precisely pinpointed in Sumbersari, Tawang Argo Village, Karangploso, in the district of Malang, located at coordinates 7.824° South Latitude and 112.578° East Longitude. This area spans an elevation range of 1200 to 1800 meters above sea level. The UB Forest area is not only prominent as the central locus for this machine learning-centered research but is also renowned for its rich ecological diversity and extensive assortment of agroforestry practices.

The UB Forest, predominantly characterized by native pine species and interspersed coffee plantations, serves as an ideal model for studying coffee-pine agroforestry systems. The forest experiences a range of climatic conditions from temperate to subtropical, attributable to its varied elevation and geographical positioning. Additionally, the soil in the UB Forest varies significantly, ranging from fertile loamy to sandy textures, offering a unique platform to explore soil-plant dynamics in agroforestry environments. The distinct features of the UB Forest, encompassing its vegetation, climatic conditions, and soil types, render it an exemplary natural laboratory. This environment is particularly conducive to the deployment and evaluation of machine learning algorithms. These algorithms are focused on filling gaps in microclimatic data and assessing their effects on the productivity and sustainability of agroforestry systems. Therefore, the research approach is fundamentally aimed at exploiting the diverse characteristics of the UB Forest. This involves developing an in-depth understanding of the microclimates within agroforestry systems and identifying contributing factors through the use of sophisticated machine-learning techniques.

2.2. Data Collection

During this study, an extensive data collection process was implemented from April 2019 to March 2020 in the Coffee-Pine agroforestry system of the UB Forest. This data included vital environmental metrics such as humidity, temperature, and sunlight intensity. Humidity levels were accurately measured using sophisticated sensors, yielding important data on soil moisture content and availability at various depths. In parallel, the combination of humidity and temperature was continually tracked using climatic instruments, which were strategically placed throughout various microclimatic areas within the forest. The temporal span of the data was instrumental in capturing the seasonal fluctuations and patterns, significantly enhancing the accuracy and pertinence of the research. This approach was crucial for a deeper understanding and effective management of the unique ecological aspects of the UB Forest.

In this research, a meticulous data collection methodology was employed, utilizing advanced instruments precisely deployed within the coffee-pine agroforestry areas of the UB Forest. These instruments included the HOBO Solar Radiation Shield paired with the Lascar Electronic Temperature & Humidity USB Data Logger, the Odyssey Soil Moisture Sensor Probe Offsets, and the MX2202 Hobo MX Temperature + Light Intensity sensor. These tools were instrumental in accurately capturing the intricate environmental conditions prevalent within the agroforestry system. To maintain the integrity and continuous flow of data, these instruments were configured to internally store the collected information. Field operators were responsible for manually downloading the data every month. This process entailed connecting each instrument to a laptop via a serial USB port, ensuring a secure and efficient transfer of data for further analysis. The arrangement and deployment of these instruments for collecting microclimate data in the UB Forest are illustrated in Fig. 1.

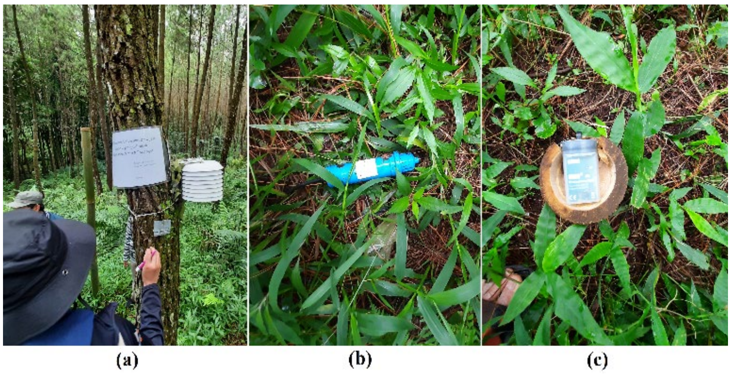

Fig. 1. (a) HOBO Solar Radiation Shield Accompanied by Lascar Electronic (Temperature and Humidity Measurement Device). (b) Soil Moisture Sensor Probe Offsets by Odyssey. (c) MX2202 Hobo MX Instrument for Measuring Temperature and Light Intensity

The instruments were deployed across various zone plots within the UB Forest, namely the BAU (Business as Usual), LC (Low Complexity), MC (Medium Complexity), and HC (High Complexity) plots. These plots were selected to represent a gradient of agroforestry system complexities and management practices, thereby providing a comprehensive dataset that reflected the diverse environmental interactions within these systems. This strategic placement of instruments across different plots was integral to capturing a wide array of microclimatic conditions and understanding their impacts on the coffee-pine agroforestry system. The collected data serves as a foundational element for applying the Data Imputation Method, aimed at elucidating the complex relationships between environmental parameters and agroforestry productivity. UB Forest area and zone plots are shown in Fig. 2.

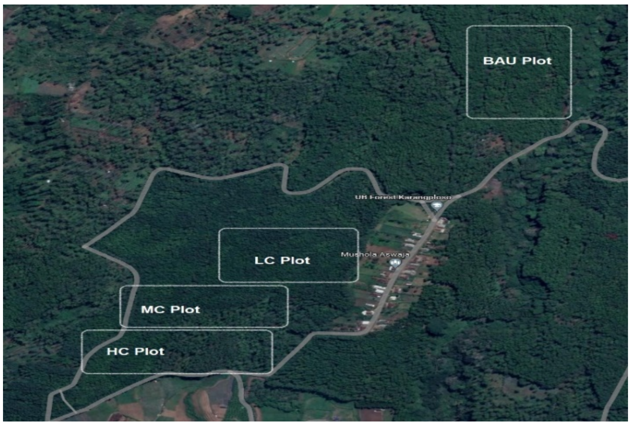

Fig. 2. UB Forest area and zone plots

2.3. Data Imputation Method

Machine learning models such as K-nearest neighbors (KNN), linear regression, and artificial neural networks are crucial in data imputation methodologies. Interpolation methods, including machine learning-based algorithms, play a significant role in filling missing data points, especially in scenarios such as daily climate data and microclimate data in agroforestry [31]. Additionally, the impact of missing data and imputation methods on the analysis of activity patterns underscores the importance of accurate imputation techniques [32]. Furthermore, the use of artificial neural networks for missing feature reconstruction highlights the relevance of advanced techniques in imputation [33]. Moreover, neural models have been employed for the imputation of missing ozone data, demonstrating the applicability of machine learning in addressing missing data in various domains [34].

The research offers a comprehensive exploration into the intricacies of handling missing microclimate data in the realm of agroforestry. The paper commences with an introduction, underscoring the importance of microclimate data in coffee-pine agroforestry and the prevailing challenges associated with data gaps. It progresses into an in-depth analysis of these challenges, particularly focusing on their impact on agroforestry research, supplemented by case research in the coffee-pine context. A pivotal section of this research is the methodology, which introduces the general approaches to data imputation, with a specific emphasis on machine learning techniques like Interpolation, Shifted Interpolation, K-Nearest Neighbors (KNN), and Linear Regression. This segment not only details each method but also evaluates its strengths and limitations in an agroforestry setting. Subsequently, the paper presents a series of case studies and experimental results that apply these methods to coffee-pine agroforestry microclimate data, offering a comparative analysis of their effectiveness. The discussion section in the paper interprets these results, linking the methodologies to their broader implications in agroforestry and environmental research. The conclusion summarizes the primary findings and suggests potential directions for future research, aiming to bridge current knowledge gaps and guide forthcoming studies.

The Interpolation Algorithm is a methodical approach used in estimating unknown values within a certain range, based on known data points. The process begins with the determination of four key values: x0, y0, x1, and y1 [35]. These values represent two known points on a line, with x0 and x1 as the independent variables and y0 and y1 as the dependent variables. Initially, the algorithm checks if x0 equals x1 [36]. If they are equal, the process is restarted since the function's value is undefined under this condition. If x0 and x1 are different, the next step involves entering a new value for x, the point at which interpolation is to be performed [37]. The algorithm then verifies whether the new x value lies within the minimum and maximum range of x0 and x1. If x does not fall within this range, a different x value is selected [38]. Once an appropriate x value is chosen, the algorithm calculates the interpolated value P. This is done using the equation (1) [39].

$$P = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} \tag{1}$$

An additional check is performed to see if y0 equals y1. If they are the same, the interpolated value P will be equal to y0. Finally, the result y=P is recorded, providing the estimated value at the chosen point x [40]. This algorithm is particularly useful in various fields for estimating values within a known range, aiding in data analysis and predictions where exact values are not available. The basic interpolation is suited for estimating values in a straightforward, linear manner, shifted interpolation is designed to handle more complex relationships, especially in time series data, by considering the time-lagged correlations. Shifted interpolation can be more sophisticated and is typically used when the data shows periodicity or patterns that are not immediately adjacent but occur at consistent intervals [41].

The KNN algorithm is a classification method that operates by identifying the closest training data to an object and using this proximity to determine the object's classification. In this algorithm, training data are mapped onto a multi-dimensional space, where each dimension corresponds to a distinct attribute of the data [42]. This space is divided into segments based on the classification of the training data. In this multi-dimensional space, a point is designated as belonging to class 'c' if class 'c' represents the most common classification among the 'k' nearest neighbors of that point [43]. The proximity of these neighbors is typically determined by the Euclidean distance, calculated using a specified equation (2) [44].

$$Distance = \sqrt{\sum_{i=1}^{n}(x1i - x2i)^2} \tag{2}$$

During the learning phase, the KNN algorithm stores the feature vectors and classifications of the training data. During the classification phase, the same features are computed for the test data, whose classification remains undetermined [45]. The distances between this new vector and all vectors in the training data are calculated, followed by the selection of the K nearest ones. The classification of the new point is then predicted based on the most prevalent classification among these selected points [46].

The optimal value of K is contingent on the data characteristics. Typically, a larger K value diminishes the impact of noise in classification, but it can also obscure the distinctions between different classifications [47]. Determining a suitable K value can be achieved using parameter optimization methods like cross-validation. A specific instance of this algorithm, where the classification is predicted based solely on the nearest training data point (that is, K equals 1), is recognized as the nearest neighbor algorithm [48]. This approach is particularly effective for data-driven decision-making in various applications, as it leverages similarity in data features for classification.

The Linear Regression algorithm encompasses a relationship between a singular dependent variable and an independent variable. In this relationship, the dependent variable (y) is affected by the independent variable (x) [49]. The relationship between the dependent and independent variables can be articulated through various forms of equations, encompassing linear, exponential, and multiple relationships [50]. The principal aim of using regression analysis is to predict the values of the dependent variable based on the values of the independent variable [51]. Linear Regression is grounded in the pattern of relationships found in historical data. Generally, the predictable variables, represented by certain variables (such as inventory levels), are influenced by the magnitude of the independent variables. The relationship that exists between the independent variable and the variable to be predicted is a function [52]. Linear Regression is characterized by the following equation (3) [53].

$$y = a + bx \tag{3}$$

In this context, y symbolizes the dependent variable, x signifies the independent variable, a represents a constant term, and b signifies the coefficient representing the response generated by the predictor.

2.4. Implementation Imputation Process

The imputation process comprises several steps. Initially, the process involves reading the data intended for use and segmenting the data range for each imputation scenario. For the time frame of 6 hours, the data range is set at 1 day; for 1 day, the range is extended to 5 days; for 1 week, the range becomes 1 month, and for 1 month, the range extends to 3 months. Subsequently, the identified data for imputation is replaced with NaN values, initiating the imputation process. For Interpolation, the data undergoes direct imputation utilizing interpolation methods. In Shifted Interpolation, the data is shifted by an amount 'n' before undergoing the interpolation process. In the case of KNN imputation, the available data undergoes training with the KNN algorithm, and subsequently, the empty data is filled using the trained KNN imputer. In linear regression, the data is divided into training and testing sets. The training set comprises data that does not contain NaN (Not a Number) values, while the testing set includes data that does contain NaN values. The linear regression model is first trained using the training set. Subsequently, predictions are made for the testing set. These predicted values are then used to replace the NaN values in the original data. Post-imputation, the effectiveness of the imputation techniques is assessed through an evaluation using metrics like MAE, MSE, RMSE, and MAPE. This involves a comparison between the original data and the imputed data to determine the accuracy and reliability of the imputation methods.

2. Results and Discussion

3.1. Datasets

The study is centered around two comprehensive datasets, which are pivotal to the research objectives. The first dataset encompasses a detailed compilation of light intensity measurements within the coffee-pine agroforestry ecosystem. These measurements are of paramount importance for evaluating the photosynthetic activity and growth conditions of various plant species within this habitat. The second dataset is expected to encompass a thorough record of air temperature and humidity, factors that are crucial in shaping the microclimate of the area. These parameters play a significant role in influencing plant physiology and the overall health of the ecosystem. Both datasets have been meticulously gathered and preserved, offering critical insights into the environmental dynamics of the UB Forest. The subsequent analysis of this data through the application of various machine learning models, including Interpolation, Shifted Interpolation, K-Nearest Neighbors (KNN), and Linear Regression, facilitates a detailed understanding of microclimatic patterns. This methodological approach is projected to make significant contributions to the field of data science, particularly in the imputation or restoration of missing microclimate data within the coffee-pine agroforestry system of the UB Forest.

3.2. Result Imputation

In this pivotal section of our research, we delve into the results derived from employing advanced machine learning algorithms to impute missing microclimate data in coffee-pine agroforestry landscapes within UB Forest. The research harnesses the potent combination of Big Data analytics and Machine Learning algorithms to address the critical challenge of data gaps in environmental monitoring. The crux of our research hinges on the utilization of various sophisticated imputation methods—namely Interpolation, Shifted Interpolation, K-Nearest Neighbors (KNN), and Linear Regression—each offering a unique approach to managing data gaps in microclimate datasets.

The statistical metrics MAE, MSE, RMSE, and MAPE are commonly used to evaluate the performance and accuracy of predictive models. MAE quantifies the average magnitude of errors between predicted and actual values. It offers an insight into the accuracy of the model, focusing on the size of the errors without taking into account the direction of these errors [54]. MSE computes the average of the squares of the errors. This approach emphasizes and gives greater weight to larger errors, effectively penalizing the model more heavily for significant deviations from the actual values [55]. RMSE represents the square root of the MSE. It provides an interpretable measure of the standard deviation of the residuals, thereby offering an indication of the model's performance in the same units as the response variable. This metric helps in understanding the average distance between the predicted values and the actual values [55]. MAPE calculates the percentage of the absolute errors about the actual values. This metric is especially useful for comparing the accuracy of different models across data sets with varying scales and magnitudes, as it provides a scale-independent measure of error [56]. These metrics are crucial in various fields such as energy forecasting, financial prediction, and manufacturing budgeting, as they provide a quantitative assessment of the predictive model's performance. For instance, in the context of energy forecasting, these metrics are used to evaluate the accuracy of models in predicting electricity consumption [56]. Similarly, in financial applications, such as exchange rate prediction, these metrics are employed to compare the performance of different predictive models, enabling the selection of the most accurate and reliable model for practical use [57].

These gaps pose significant hurdles in understanding and managing agroforestry ecosystems effectively. By applying advanced imputation techniques, we aim to reconstruct a comprehensive microclimatic profile of the area, which is crucial for assessing the health and sustainability of the coffee-pine agroforestry system. This section not only presents the outcomes of our data imputation efforts but also critically analyzes the implications of these findings in the broader context of agroforestry management and environmental conservation. Through this discussion, we aim to contribute to the ongoing discourse in agroforestry research and demonstrate the application of cutting-edge technology in ecological data recovery and analysis.

3.3. Result imputation in a time frame daily

Below is the graph of humidity imputation results using the Interpolation, Shifted Interpolation, KNN, and Linear Regression methods in a time frame daily as shown in Fig. 3.

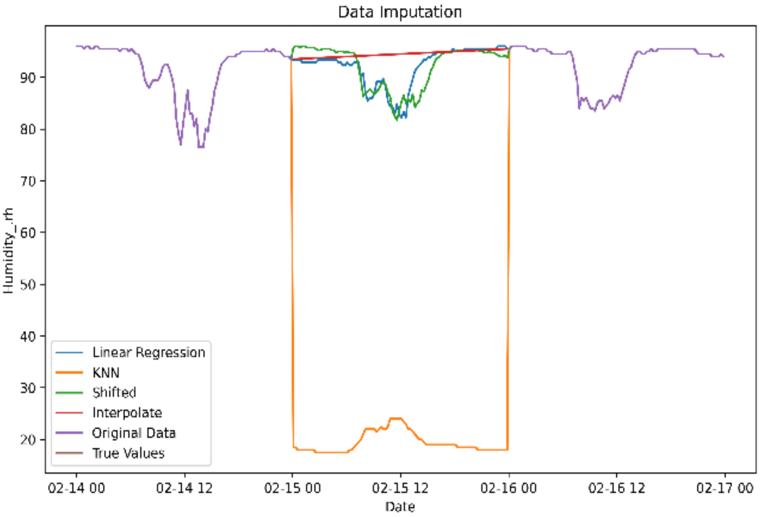

Fig. 3. Graph of humidity imputation results in the time frame daily

Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for humidity in a time frame daily can be seen in Table 1.

Table 1. Evaluation metrics imputation for humidity in time frame daily

Methods	MAE	MSE	RMSE	MAPE
Interpolation	0.86133	6.16536	2.48301	0.00949
Shifted	4.2943	39.88129	6.31516	0.04793
KNN	24.227	1771.074	42.08413	0.43435
Linear Regresion	0.055757	0.01717	0.13104	0.00061

Linear Regression demonstrated remarkable precision, with the lowest MAE (0.055757), MSE (0.01717), RMSE (0.13104), and MAPE (0.00061). These results suggest a high degree of accuracy and minimal deviation from actual values, as MAE and RMSE are direct measures of error magnitude, with lower values indicating better model performance [58]. MAPE, a relative error metric, further underscores the model's accuracy in percentage terms, which is particularly useful for comparing models across different scales [59].

Interpolation followed as the second most effective method, with a relatively low MAE (0.861330), MSE (6.16536), and RMSE (2.48301), but a higher MAPE (0.00949) than Linear Regression. This suggests a reasonable approximation of missing data, albeit less precise than Linear Regression.

Conversely, Shifted Interpolation and KNN exhibited significantly higher errors across all metrics. The KNN method, in particular, showed the highest MAE (24.227), MSE (1771.07378), RMSE (42.08413), and MAPE (0.43435), indicating substantial deviation from actual values. These larger errors could be attributed to the method's sensitivity to the dataset's specific characteristics or potential overfitting [59].

For the imputation of missing microclimate data for humidity in a time frame daily, Linear Regression emerged as the most accurate and reliable method for imputing missing daily humidity data in this research. Its superior performance is evident across all evaluation metrics, marking it as the best choice for such imputations in similar micro-climate studies.

Below is the graph of temperature imputation results using the Interpolation, Shifted Interpolation, KNN, and Linear Regression methods in a time frame daily as shown in Fig. 4.

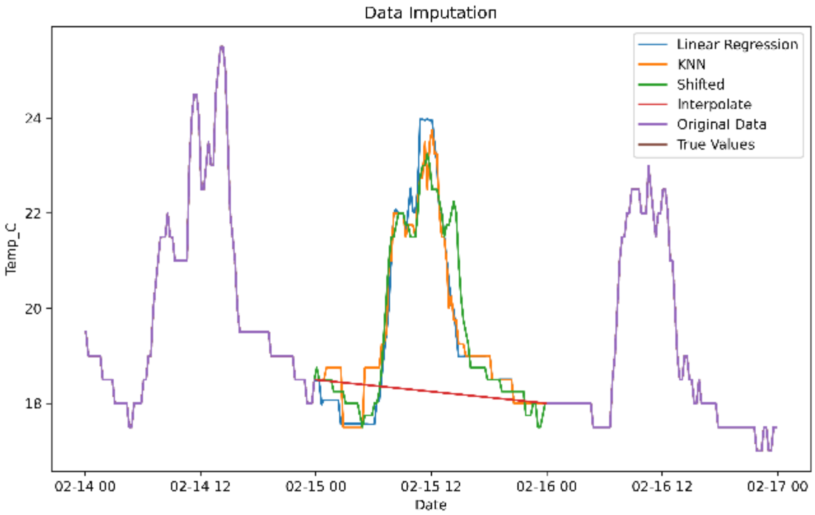

Fig. 4. Graph of temperature imputation results in the time frame daily

Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for temperature in a time frame daily can be seen in Table 2.

Table 2. Evaluation metrics imputation for Temperature in time frame daily

Methods	MAE	MSE	RMSE	MAPE
Interpolation	0.50943	1.73778	1.31825	0.0257
Shifted	2.105	7.80968	2.79458	0.105
KNN	0.092882	0.08312	0.2883	0.00471
Linear Regresion	0.010152	0.00056	0.02359	0.00054

The KNN method exhibited exceptional precision in imputing temperature data, as reflected by the lowest MAE (0.092882), MSE (0.08312), RMSE (0.28830), and MAPE (0.00471). These results suggest a high degree of accuracy with minimal errors. MAE and RMSE are direct measures of error magnitude, where lower values indicate better model performance, and the KNN method's low values point to its effectiveness in accurately predicting temperature data [59]. Furthermore, the low MAPE value signifies a minimal percentage error, making KNN highly suitable for temperature data imputation in this context [60].

Linear Regression also performed admirably, yielding extremely low MAE (0.010152), MSE (0.00056), RMSE (0.02359), and MAPE (0.00054). These metrics indicate a high level of precision, but the slightly higher values compared to KNN suggest that for this specific dataset, KNN might be more adept at handling the nuances of temperature data imputation.

Interpolation and Shifted Interpolation methods showed comparatively higher error metrics, with Shifted Interpolation, in particular, exhibiting significantly higher MAE, MSE, RMSE, and MAPE. These higher values indicate less accuracy in imputing temperature data compared to KNN and Linear Regression. For the imputation of missing microclimate data for temperature in a time frame daily, the KNN method stands out as the most effective for daily temperature data imputation in this research, balancing accuracy with computational efficiency. Its superiority is demonstrated across all key performance metrics, establishing it as the preferred method for temperature data imputation in similar micro-climate studies.

Below is a graph of intensity results in the imputation of intensity using the Interpolation, Shifted Interpolation, KNN, and Linear Regression methods in the time frame daily shown in the Fig. 5.

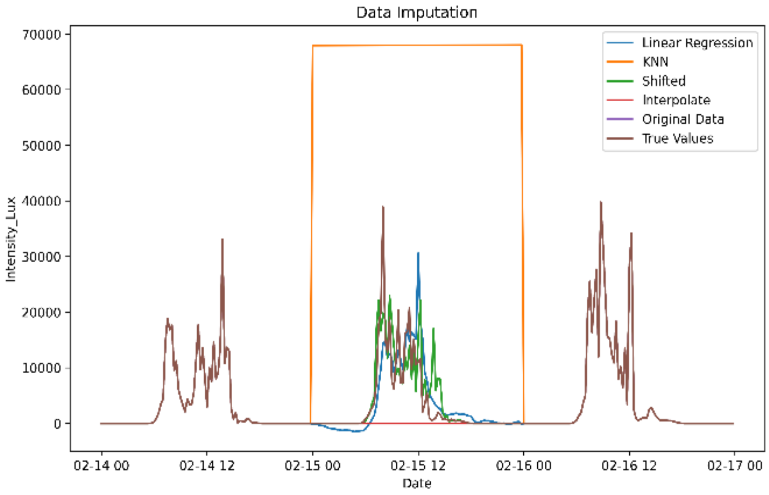

Fig. 5. Graph of intensity imputation results in the time frame daily

Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for intensity in a time frame daily can be seen in Table 3.

Table 3. Evaluation metrics imputation for Intensity in time frame daily

Methods	MAE	MSE	RMSE	MAPE
Interpolation	1177.2	19283359	4391.282	0.35417
Shifted	6264.3	1.1E+08	10464.85	1.48641
KNN	21490	1.4E+09	37424.09	0.61016
Linear Regresion	759.33	6447309	2539.155	0.47196

Linear Regression emerged as the most effective method for imputing intensity data, as evidenced by its relatively low MAE (759.33), MSE (6,447,309), RMSE (2539.15515), and MAPE (0.47196). These metrics indicate a strong balance of accuracy and consistency in the model's predictions. The lower MAE and RMSE values suggest that Linear Regression was generally closer to the true values, with fewer and less severe errors, a critical aspect in imputation tasks [61]. The MAPE value, although not the lowest among the models, still reflects a reasonably low percentage error in predictions.

Interpolation, while not as precise as Linear Regression, showed moderate effectiveness with an MAE of 1177.20, MSE of 19,283,360, RMSE of 4391.28211, and the lowest MAPE (0.35417) among the methods. This suggests a decent level of accuracy, particularly in relative terms (percentage error), but with higher absolute errors compared to Linear Regression. Shifted Interpolation and KNN, on the other hand, exhibited significantly higher errors across all metrics. Particularly, the KNN method displayed the highest MAE (21,490.00), MSE (1,400,562,000), RMSE (37,424.08797), and a high MAPE (0.61016), indicating a substantial deviation from actual values. These elevated error levels could be indicative of the methods' limitations in capturing the complexities of daily intensity data in this specific context.

For the imputation of missing microclimate data for intensity in a time frame daily, Linear Regression was the most proficient method for imputing daily intensity data in this research. Its performance, as reflected by the evaluated metrics, underscores its suitability for accurate and reliable imputation in similar micro-climate data analysis scenarios.

3.4. Result imputation in a time frame weekly

Below is the graph of humidity imputation results using the Interpolation, Shifted Interpolation, KNN, and Linear Regression methods in a time frame weekly as shown in Fig. 6.

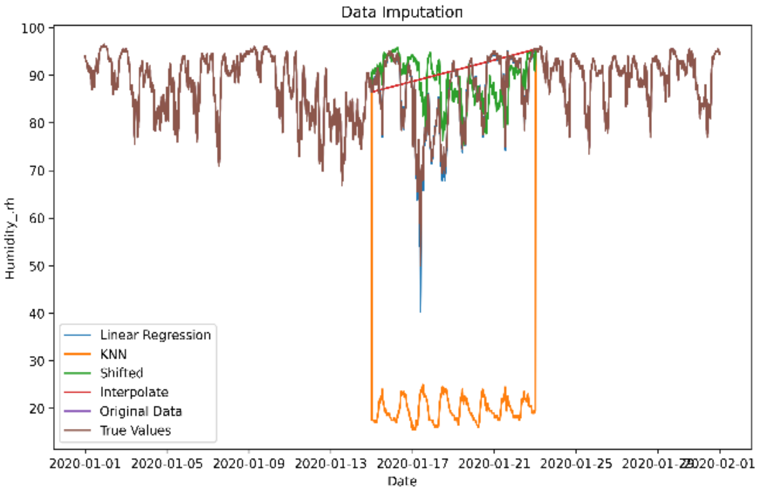

Fig. 6. Graph of humidity imputation results in the time frame weekly

Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for humidity in a time frame weekly can be seen in Table 4.

Table 4. Evaluation metrics imputation for Humidity in time frame weekly

Methods	MAE	MSE	RMSE	MAPE
Interpolation	1.6862	20.69972	4.54969	0.01989
Shifted	6.2204	63.31979	7.95737	0.07192
KNN	17.155	1161.234	34.07689	0.32259
Linear Regresion	0.085474	0.12577	0.35465	0.00114

Linear Regression demonstrated exceptional accuracy in predicting weekly humidity levels, as evidenced by the lowest MAE (0.085474), MSE (0.12577), RMSE (0.35465), and MAPE (0.00114) among the tested methods. These results suggest a high level of precision and minimal deviation from actual humidity values, crucial in micro-climate data analysis. Lower MAE and RMSE values indicate more accurate predictions with fewer errors, making Linear Regression particularly effective for imputing missing data in this context [62]. Additionally, the remarkably low MAPE underscores the model's efficiency in relative error terms, further affirming its suitability for this task [59].

Interpolation also performed well, with a relatively low MAE (1.686200), MSE (20.69972), RMSE (4.54969), and a moderate MAPE (0.01989). Although not as accurate as Linear Regression, these figures suggest that Interpolation is a reliable method for imputing weekly humidity data, offering a good balance between simplicity and accuracy.

In contrast, Shifted Interpolation and KNN displayed significantly higher errors across all metrics. The KNN method, in particular, showed the highest MAE (17.155), MSE (1161.23412), RMSE (34.07689), and MAPE (0.32259), indicating a considerable deviation from the actual humidity values. These larger errors could be attributed to the methods' inability to capture the variability and patterns in weekly humidity data effectively [59].

In the imputation of missing microclimate data for humidity in a time frame weekly, Linear Regression emerged as the superior method for imputing weekly humidity data in this research. Its performance across all evaluation metrics highlights its robustness and reliability, making it the preferred choice for similar imputation tasks in micro-climate studies.

Below is the graph of temperature imputation results using the Interpolation, Shifted Interpolation, KNN, and Linear Regression methods in a time frame weekly as shown in Fig. 7.

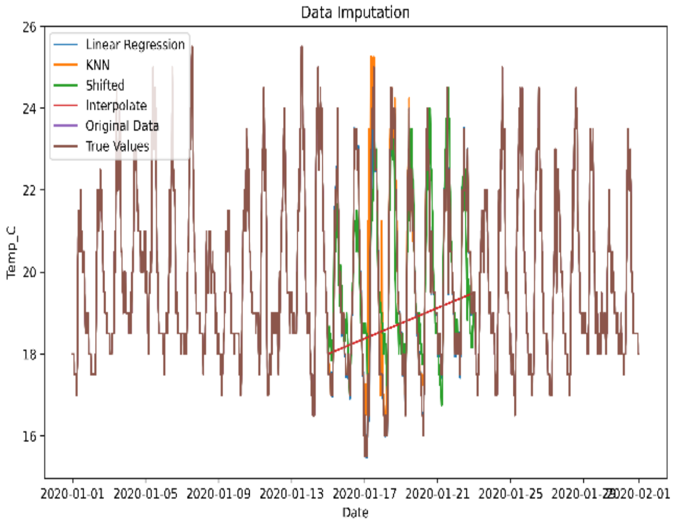

Fig. 7. Graph of temperature imputation results in the time frame weekly

Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for temperature in a time frame weekly can be seen in Table 5.

Table 5. Evaluation metrics imputation for Temperature in time frame weekly

Methods	MAE	MSE	RMSE	MAPE
Interpolation	0.50867	1.5998	1.26483	0.02572
Shifted	2.1226	7.26125	2.69467	0.10531
KNN	0.056704	0.13187	0.36314	0.00297
Linear Regresion	0.016263	0.00468	0.06841	0.00079

The KNN method displayed outstanding precision in estimating weekly temperature data, evidenced by the lowest MAE (0.056704), MSE (0.13187), RMSE (0.36314), and MAPE (0.00297). These metrics indicate an exceptional level of accuracy, with minimal deviation from the actual values. Lower MAE and RMSE values are particularly significant as they represent direct measures of error magnitude, with smaller values suggesting higher accuracy in predictions. The low MAPE value also points to minimal relative error, further emphasizing the efficacy of the KNN method in this context [59]. Linear Regression also performed well, with very low MAE (0.016263), MSE (0.00468), RMSE (0.06841), and MAPE (0.00079). Although slightly outperformed by the KNN method, these figures underscore the high precision of Linear Regression in imputing weekly temperature data.

On the other hand, Interpolation and Shifted Interpolation methods showed comparatively higher error metrics, with Shifted Interpolation particularly displaying significantly higher MAE, MSE, RMSE, and MAPE. These elevated error levels suggest less accuracy and reliability in these methods for predicting weekly temperature data, potentially due to their less sophisticated handling of temporal dynamics in the data [59]. The imputation of missing microclimate data for temperature in a time frame weekly, the KNN method emerges as the most effective for imputing weekly temperature data in this research. Its superior performance across all evaluated metrics highlights its robustness and suitability for such tasks, particularly in the context of micro-climate data analysis in agroforestry settings. Fig. 8 is the graph of intensity imputation results using the Interpolation, Shifted Interpolation, KNN, and Linear Regression methods in a time frame weekly.

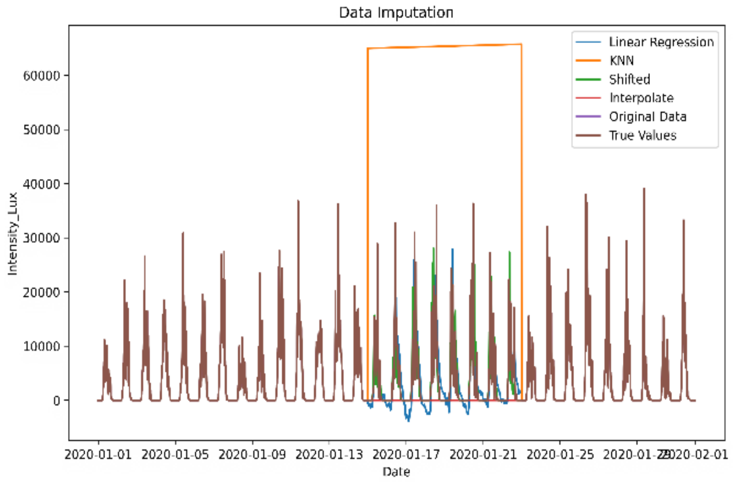

Fig. 8. Graph of Intensity imputation results in the time frame weekly

Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for intensity in a time frame weekly can be seen in Table 6. Linear Regression exhibited the most accurate performance for weekly intensity data imputation, as indicated by its lowest MAE (642.12), MSE (3,712,463), RMSE (1926.77533), and moderately low MAPE (0.34605). These metrics suggest a high level of precision in the model's predictions, with minimal deviation from the actual values. Lower values in MAE and RMSE are particularly significant as they represent a more accurate prediction with fewer errors, a crucial factor in data imputation tasks [59]. The MAPE value, while not the lowest, still signifies a reasonable accuracy in relative terms, important for understanding the model's performance in percentage error terms [58].

Interpolation also showed reasonable effectiveness with an MAE of 918.13, MSE of 12,114,410, RMSE of 3480.57659, and MAPE of 0.28360, indicating a fair level of accuracy in imputing weekly intensity data, though not as precise as Linear Regression. Contrastingly, Shifted Interpolation, and KNN showed significantly higher errors across all metrics. Notably, the KNN method displayed the highest MAE (15,949), MSE (994,558,800), RMSE (31,536.62559), and a high MAPE (0.46890), indicating substantial deviations from the actual values. This may be due to the methods' limitations in effectively capturing the complex patterns in weekly intensity data [59].

Table 6. Evaluation metrics imputation for Intensity in time frame weekly

Methods	MAE	MSE	RMSE	MAPE
Interpolation	918.13	12114413	3480.577	0.2836
Shifted	4979.4	61287825	7828.654	1.30729
KNN	15949	9.95E+08	31536.63	0.4689
Linear Regresion	642.12	3712463	1926.775	0.34605

The imputation of missing microclimate data for intensity in a time frame weekly, Linear Regression stood out as the most effective method for imputing weekly intensity data in this research. Its performance across all key metrics underscores its robustness and reliability, making it the preferred choice for similar imputation tasks in micro-climate studies.

3.5. Result Imputation in all across the time frame

In this research, a comprehensive data imputation method is used which is adapted to the intricacies of microclimate data analysis. At the heart of our research are four different imputation methodologies: Interpolation, Shifted Interpolation, K-Nearest Neighbors (KNN), and Linear Regression. Each of these methods presents a unique approach to addressing gaps in microclimate data sets, which are often caused by sensor malfunctions, environmental interference, or data transmission errors. The novelty of this research lies in its application at various temporal resolutions, including 6-hour, daily, weekly, and monthly time frames. This detailed analysis allows for a different understanding of the temporal dynamics in the Coffee-Pine Agroforestry ecosystem. By carefully evaluating these methods over multiple time scales, this research aims not only to identify the most effective imputation strategies for each climate parameter but also to provide insight into the temporal patterns of agroforestry microclimates. This multifaceted approach plays an important role in advancing the field of ecological data science, particularly in the context of sustainable agroforestry management and climate change adaptation. Evaluation metrics of Interpolation, Shifted Interpolation, KNN, and Linear Regression methods to the imputation of missing microclimate data for humidity, temperature, and intensity in time frame 6-hour, daily, weekly, and monthly can be seen in Table 7.

The key objective was to identify the most effective method for each variable and time frame, guided by standard evaluation metrics. The humidity Imputation for 6 Hours to Monthly Across all time frames, Linear Regression consistently outperformed the other methods. It demonstrated the lowest MAE, MSE, RMSE, and MAPE, indicating superior accuracy and reliability. This consistency is crucial, as lower MAE and RMSE values signify a model's ability to predict with fewer errors, an essential aspect in time-sensitive micro-climate data imputation [62]. The temperature Imputation for 6 Hours to Monthly in The KNN method excelled in shorter time frames (6 hours), showcasing its proficiency in handling rapid temperature fluctuations. For daily to monthly time frames, Linear Regression emerged as the most accurate, reflecting its effectiveness in capturing longer-term temperature trends [58]. The Intensity Imputation for 6 Hour to Monthly, the Linear Regression again showed its superiority, yielding the lowest error metrics across all time frames. This indicates its strong predictive power for intensity data, a key factor in comprehensive climate modeling [59].

In the theoretical Implications and Best Model Selection, The Linear Regression method stands out as the most effective, particularly in the context of humidity and temperature imputation. Its low MAE, MSE, RMSE, and MAPE signify high accuracy and reliability, crucial in micro-climate data analysis where precise predictions are vital [63]. The KNN method, despite its popularity in classification problems, shows limitations in this context, especially for humidity and intensity data imputation. This could be attributed to its sensitivity to the local data structure, which might not be optimal for the spatial-temporal nature of micro-climate data [59]. Interpolation, a basic yet often effective method, shows balanced performance across all variables, particularly in situations where data points are closely aligned in time or space [63]. Shifted Interpolation, despite its potential in handling lagged correlations, does not perform well in this dataset, possibly due to the complex interactions in micro-climate variables not aligning well with shifted patterns [64].

Table 7. Evaluation Metrics All Time Frame For Prediction

Time Frame	Methods	Humidity				Temperature				Intensity			
		MAE	*MSE*	*RMSE*	*MAPE*	*MAE*	*MSE*	*RMSE*	*MAPE*	*MAE*	*MSE*	*RMSE*	*MAPE*
6 hour	Interpolation	0.76419	3.91783	1.97935	0.00875	0.37949	0.88928	0.94302	0.0174	1268.6	8127124	2850.8111	0.22009
	Shifted	3.151	20.49609	4.52726	0.03466	1.8281	6.07812	2.46539	0.09239	5905	99575107	9978.7328	1.40495
	KNN	19.521	1316.1615	36.27894	0.35365	0.25521	0.37109	0.60917	0.01167	17813	1.1E+09	33163.462	0.48278
	Linear Regression	0.052594	0.01342	0.11584	0.0006	0.0096478	0.00045	0.02116	0.00044	2662.7	42166261	6493.5553	0.2869
Daily	Interpolation	0.86133	6.16536	2.48301	0.00949	0.50943	1.73778	1.31825	0.0257	1177.2	19283359	4391.2821	0.35417
	Shifted	4.2943	39.88129	6.31516	0.04793	2.105	7.80968	2.79458	0.105	6264.3	109512991	10464.845	1.48641
	KNN	24.227	1771.0738	42.08413	0.43435	0.092882	0.08312	0.2883	0.00471	21490	1.401E+09	37424.088	0.61016
	Linear Regression	0.055757	0.01717	0.13104	0.00061	0.010152	0.00056	0.02359	0.00054	759.33	6447308.9	2539.1552	0.47196
weekly	Interpolation	1.6862	20.69972	4.54969	0.01989	0.50867	1.5998	1.26483	0.02572	918.13	12114413	3480.5766	0.2836
	Shifted	6.2204	63.31979	7.95737	0.07192	2.1226	7.26125	2.69467	0.10531	4979.4	61287825	7828.6541	1.30729
	KNN	17.155	1161.2341	34.07689	0.32259	0.056704	0.13187	0.36314	0.00297	15949	994558753	31536.626	0.4689
	Linear Regression	0.085474	0.12577	0.35465	0.00114	0.016263	0.00468	0.06841	0.00079	642.12	3712463.2	1926.7753	0.34605
monthly	Interpolation	1.1157	8.26569	2.87501	0.01231	0.54434	1.76695	1.32927	0.02777	1142.4	16908060	4111.9412	0.34112
	Shifted	5.862	60.95852	7.80759	0.06708	2.414	8.89922	2.98316	0.1206	5342.5	69612692	8343.4221	1.47745
	KNN	22.913	1657.0992	40.70748	0.41308	0.0070423	0.00434	0.06586	0.00042	20528	1.335E+09	36541.61	0.58151
	Linear Regresion	0.074267	0.02371	0.15398	0.00081	0.013607	0.0008	0.02832	0.00071	657	4279478.4	2068.69	0.42543

The research found Linear Regression to be the most effective method for imputing humidity and intensity data across various time frames. For temperature data, KNN proved to be more suitable for shorter time frames, while Linear Regression was preferable for longer durations. These findings offer valuable insights for future micro-climate data imputation tasks, highlighting the importance of selecting appropriate models based on specific variables and time frames. In short, imputation of microclimate lost data for all parameters in different imputation methods, the Linear Regression method is identified as the best model for imputing lost microclimate data in the studied context. Its superiority in accuracy across multiple metrics for humidity, temperature, and intensity makes it a robust choice for such applications. Future research could explore the integration of Linear Regression with other techniques to further enhance imputation accuracy in complex agroforestry micro-climates.

One of the primary challenges in implementing these machine learning methods is the requirement for substantial computational resources, which may not be readily available in all agroforestry management settings. The complexity of these algorithms also necessitates a certain level of expertise in data science, which could be a barrier for practitioners without a technical background. Furthermore, the accuracy of these imputation methods is highly dependent on the quality of the available data. If the existing datasets are sparse or biased, the imputed values could inadvertently introduce errors, leading to misguided decisions in agroforestry management. Additionally, the methods may have varying levels of efficacy depending on the specific environmental context, the scale of application, and the type of microclimate data being analyzed. To integrate these methods into existing management practices, it is suggested that practitioners collaborate with data scientists to create streamlined tools that automate much of the complex processes involved. Building user-friendly interfaces and providing training on interpreting the outputs of these models can also facilitate their adoption. Acknowledging the limitations of this research, there is a need for future studies to explore the scalability of these methods in diverse agroforestry systems and to assess the long-term impacts of management decisions informed by imputed data. Further research could also investigate the integration of additional variables that could affect microclimate data, such as topographical features or specific agricultural practices, to refine the imputation models.

3. Conclusion

Linear Regression consistently emerged as the most effective method for imputing humidity and intensity data across all time frames (6 hours, daily, weekly, monthly). Its superior performance is indicated by the lowest MAE, MSE, RMSE, and MAPE values, demonstrating high accuracy and reliability. This underscores the robustness of Linear Regression in handling diverse micro-climate data sets over varying temporal scales. For temperature data, the KNN method excelled in shorter time frames (6 hours), highlighting its capability to capture rapid temperature changes. Conversely, Linear Regression proved more effective in longer time frames (daily, weekly, monthly), indicating its strength in modeling long-term temperature trends. This suggests that the choice of imputation method should be tailored to the specific characteristics of the data, particularly the nature and frequency of the variable in question. Accurate imputation of micro-climate data is crucial for environmental management and agricultural planning, especially in sensitive ecosystems like agroforestry landscapes. The findings provide essential insights for practitioners and researchers in these fields, offering a guideline on the most suitable methods for data imputation based on their specific requirements and data characteristics. For temperature data, which often exhibits more rapid fluctuations, the K-Nearest Neighbors (KNN) method proves superior in short-term scenarios (6 hours), while Linear Regression excels over longer periods, reflecting its ability to capture and model sustained temperature trends effectively. These findings suggest a nuanced approach to data imputation: the selection of an imputation technique should be customized to the specific nature of the data and the time scale of interest. Such tailored applications are critical for ensuring the precision of environmental and agricultural models, which, in turn, are vital for the sound management of sensitive agroforestry ecosystems. The implications of this research are profound, offering practitioners and researchers a guided methodology for enhancing the quality of microclimate data analysis. This advancement supports more informed decision-making in environmental management and fosters improved agricultural planning, contributing significantly to the sustainability and resilience of agroforestry landscapes.

Acknowledgment

The authors extend their sincere gratitude to the UK Centre for Ecology & Hydrology (UKCEH), located at the Lancaster Environment Centre, Lancaster, United Kingdom, for their valuable collaboration in research and the provision of research equipment. Additionally, the authors are appreciative of the financial support received for this research, generously provided by the Strengthening Research Ecosystem of Professor Grant Scheme 2023, under contract number 1759.1.15/UN10.C20/2023. This support was facilitated by the Institute of Research and Community Services at Brawijaya University (LPPM-UB), Malang, Indonesia. Gratitude is also extended to the management of UB Forest and the local farmers, whose permission was instrumental in conducting the research activities. The authors also wish to acknowledge the AI Research Center at UB for their contributions.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper

Funding statement. The Strengthening Research Ecosystem of Professor Grant Scheme 2023, bearing the contract number 1759.1.15/UN10.C20/2023, is a notable initiative under the auspices of the Institute of Research and Community Services at Brawijaya University (LPPM-UB) in Malang, Indonesia. This grant scheme plays a crucial role in supporting and enhancing research endeavors and facilitating academic and scientific advancements within the university and the broader research community.

Conflict of interest. The authors have stated that there are no conflicts of interest regarding this research.

Additional information. No additional information is available for this paper.

References

[1] A. Panozzo *et al.*, “Impact of Olive Trees on the Microclimatic and Edaphic Environment of the Understorey Durum Wheat in an Alley Orchard of the Mediterranean Area,” *Agronomy*, vol. 12, no. 2, p. 527, Feb. 2022, doi: 10.3390/agronomy12020527.

[2] D. Purnomo, M. Theresia Sri Budiastuti, and D. Setyaningrum, “The role of soybean agroforestry in mitigating climate change in Indonesia,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1016, no. 1, p. 012024, Apr. 2022, doi: 10.1088/1755-1315/1016/1/012024.

[3] B. S. Iskandar, J. Iskandar, R. Partasasmita, and R. L. Alfian, “Planting coffee and take care of forest: A case study on coffee cultivation in the forest carried out among people of Palintang, Highland of Bandung, West Java, Indonesia,” *Biodiversitas J. Biol. Divers.*, vol. 19, no. 6, pp. 2183–2195, Oct. 2018, doi: 10.13057/biodiv/d190626.

[4] A. I. de Castro, Y. Shi, J. M. Maja, and J. M. Peña, “UAVs for Vegetation Monitoring: Overview and Recent Scientific Contributions,” *Remote Sens.*, vol. 13, no. 11, p. 2139, May 2021, doi: 10.3390/rs13112139.

[5] E. D. Cahyono *et al.*, “Agroforestry Innovation through Planned Farmer Behavior: Trimming in Pine–Coffee Systems,” *Land*, vol. 9, no. 10, p. 363, Sep. 2020, doi: 10.3390/land9100363.

[6] H. Douville *et al.*, “Water remains a blind spot in climate change policies,” *PLOS Water*, vol. 1, no. 12, p. e0000058, Dec. 2022, doi: 10.1371/journal.pwat.0000058.

[7] A. F. S. Pino, Z. Y. D. Espinosa, and E. V. R. Cabrera, “Characterization of the Rhizosphere Bacterial Microbiome and Coffee Bean Fermentation in the Castillo-Tambo and Bourbon Varieties in the Popayán-Colombia Plateau,” *BMC Plant Biol.*, vol. 23, no. 1, p. 217, Apr. 2023, doi: 10.1186/s12870-023-04182-2.

[8] L. Guellouz and F. Khayat, “A data completion method for identifying pollution intrusion in aquifers,” *Sci. Rep.*, vol. 12, no. 1, p. 16200, Sep. 2022, doi: 10.1038/s41598-022-20131-9.

[9] J. N. Cape, R. I. Smith, and D. Leaver, "Missing data in spatiotemporal datasets: the <scp>UK</scp> rainfall chemistry network," *Geosci. Data J.*, vol. 2, no. 1, pp. 25–30, Jul. 2015, doi: 10.1002/gdj3.24.

[10] K. M. Fouad, M. M. Ismail, A. T. Azar, and M. M. Arafa, "Advanced methods for missing values imputation based on similarity learning," *PeerJ Comput. Sci.*, vol. 7, p. e619, Jul. 2021, doi: 10.7717/peerj-cs.619.

[11] T. N. Fatyanosa, N. A. Firdausanti, L. F. J. Soto, I. M. dos Santos, P. H. N. Prayoga, and M. Aritsugi, "Conducting Vessel Data Imputation Method Selection Based on Dataset Characteristics," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1198, no. 1, p. 012017, Jun. 2023, doi: 10.1088/1755-1315/1198/1/012017.

[12] L. A. Menéndez García *et al.*, "A Method of Pruning and Random Replacing of Known Values for Comparing Missing Data Imputation Models for Incomplete Air Quality Time Series," *Appl. Sci.*, vol. 12, no. 13, p. 6465, Jun. 2022, doi: 10.3390/app12136465.

[13] H. Voß *et al.*, "HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values," *Nat. Commun.*, vol. 13, no. 1, p. 3523, Jun. 2022, doi: 10.1038/s41467-022-31007-x.

[14] S. Batra, R. Khurana, M. Z. Khan, W. Boulila, A. Koubaa, and P. Srivastava, "A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records," *Entropy*, vol. 24, no. 4, p. 533, Apr. 2022, doi: 10.3390/e24040533.

[15] J. S. Sunny *et al.*, "Anomaly Detection Framework for Wearables Data: A Perspective Review on Data Concepts, Data Analysis Algorithms and Prospects," *Sensors*, vol. 22, no. 3, p. 756, Jan. 2022, doi: 10.3390/s22030756.

[16] L. Zhang, "A Pattern-Recognition-Based Ensemble Data Imputation Framework for Sensors from Building Energy Systems," *Sensors*, vol. 20, no. 20, p. 5947, Oct. 2020, doi: 10.3390/s20205947.

[17] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, vol. 135, p. 109864, Jun. 2020, doi: 10.1016/j.chaos.2020.109864.

[18] L. Erhan, M. Di Mauro, A. Anjum, O. Bagdasar, W. Song, and A. Liotta, "Embedded Data Imputation for Environmental Intelligent Sensing: A Case Study," *Sensors*, vol. 21, no. 23, p. 7774, Nov. 2021, doi: 10.3390/s21237774.

[19] Z. L. Wang, "Triboelectric Nanogenerator (TENG)—Sparking an Energy and Sensor Revolution," *Adv. Energy Mater.*, vol. 10, no. 17, p. 2000137, May 2020, doi: 10.1002/aenm.202000137.

[20] Z. Liu, C. Peng, T. Work, J.-N. Candau, A. DesRochers, and D. Kneeshaw, "Application of machine-learning methods in forest ecology: recent progress and future challenges," *Environ. Rev.*, vol. 26, no. 4, pp. 339–350, Dec. 2018, doi: 10.1139/er-2018-0034.

[21] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9000–9008, doi: 10.1109/CVPR.2018.00938.

[22] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-Aware Video Frame Interpolation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, vol. 2019-June, pp. 3698–3707, doi: 10.1109/CVPR.2019.00382.

[23] X. Zhang and X. Wu, "Image Interpolation by Adaptive 2-D Autoregressive Modeling and Soft-Decision Estimation," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 887–896, Jun. 2008, doi: 10.1109/TIP.2008.924279.

[24] F. X. Timmes and F. D. Swesty, "The Accuracy, Consistency, and Speed of an Electron-Positron Equation of State Based on Table Interpolation of the Helmholtz Free Energy," *Astrophys. J. Suppl. Ser.*, vol. 126, no. 2, pp. 501–516, Feb. 2000, doi: 10.1086/313304.

[25] T.-L. Cheng, Y.-Y. Lin, X. Lu, and R. Singh, "On Partially Linear Single-Index Models with Missing Response and Error-in-Variable Predictors," *J. Stat. Theory Appl.*, vol. 18, no. 1, p. 46, Apr. 2019, doi: 10.2991/jsta.d.190306.006.

[26] J. Poulos and R. Valle, "Missing Data Imputation for Supervised Learning," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 186–196, Apr. 2018, doi: 10.1080/08839514.2018.1448143.

[27] P. W. Bernhardt, “Model validation and influence diagnostics for regression models with missing covariates,” *Stat. Med.*, vol. 37, no. 8, pp. 1325–1342, Apr. 2018, doi: 10.1002/sim.7584.

[28] A. B. Nassif, D. Ho, and L. F. Capretz, “Towards an early software estimation using log-linear regression and a multilayer perceptron model,” *J. Syst. Softw.*, vol. 86, no. 1, pp. 144–160, Jan. 2013, doi: 10.1016/j.jss.2012.07.050.

[29] H. J. Einhorn, D. N. Kleinmuntz, and B. Kleinmuntz, “Linear regression and process-tracing models of judgment.,” *Psychol. Rev.*, vol. 86, no. 5, pp. 465–485, Sep. 1979, doi: 10.1037/0033-295X.86.5.465.

[30] G. Dudek, “Pattern-based local linear regression models for short-term load forecasting,” *Electr. Power Syst. Res.*, vol. 130, pp. 139–147, Jan. 2016, doi: 10.1016/j.epsr.2015.09.001.

[31] S. Ren *et al.*, “Machine Learning Based Algorithms to Impute PaO 2 from SpO2 Values and Development of an Online Calculator,” *Res. Sq.*, p. 16, Nov. 2021, doi: 10.21203/rs.3.rs-1053360/v1.

[32] L. Weed, R. Lok, D. Chawra, and J. Zeitzer, “The Impact of Missing Data and Imputation Methods on the Analysis of 24-Hour Activity Patterns,” *Clocks & Sleep*, vol. 4, no. 4, pp. 497–507, Sep. 2022, doi: 10.3390/clockssleep4040039.

[33] M. Friedjungová, M. Jiřina, and D. Vašata, “Missing Features Reconstruction and Its Impact on Classification Accuracy,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11538 LNCS, Springer Verlag, 2019, pp. 207–220, doi: 10.1007/978-3-030-22744-9_16.

[34] Á. Arroyo, Á. Herrero, V. Tricio, E. Corchado, and M. Woźniak, “Neural Models for Imputation of Missing Ozone Data in Air-Quality Datasets,” *Complexity*, vol. 2018, pp. 1–14, 2018, doi: 10.1155/2018/7238015.

[35] C. Kontos and D. Karlis, “Football analytics based on player tracking data using interpolation techniques for the prediction of missing coordinates,” *Stat. Appl. - Ital. J. Appl. Stat.*, vol. 35, no. 2, p. 19, May 2023. [Online]. Available at: https://www.sa-ijas.org/ojs/index.php/sa-ijas/article/view/202.

[36] H. Späth, *Mathematical algorithms for linear regression*. Academic Press, pp. 17-192, 1992, doi: 10.1016/B978-0-12-656460-0.50008-2.

[37] P. Saeipourdizaj, P. Sarbakhsh, and A. Gholampour, “Application of imputation methods for missing values of PM 10 and O 3 data: Interpolation, moving average and K-nearest neighbor methods,” *Environ. Heal. Eng. Manag.*, vol. 8, no. 3, pp. 215–226, Sep. 2021, doi: 10.34172/EHEM.2021.25.

[38] Y. Sun, T. Yang, and Z. Liu, “A whale optimization algorithm based on quadratic interpolation for high-dimensional global optimization problems,” *Appl. Soft Comput.*, vol. 85, p. 105744, Dec. 2019, doi: 10.1016/j.asoc.2019.105744.

[39] K. Dashdondov, K. Jo, and M.-H. Kim, “Linear interpolation and Machine Learning Methods for Gas Leakage Prediction Base on Multi-source Data Integration,” *J. Korea Converg. Soc.*, vol. 13, no. 3, pp. 33–41, 2022, [Online]. Available at: https://koreascience.kr/article/JAKO202210459406089.pdf.

[40] Y. Dong, Z. Fu, Y. Peng, Y. Zheng, H. Yan, and X. Li, “Precision fertilization method of field crops based on the Wavelet-BP neural network in China,” *J. Clean. Prod.*, vol. 246, p. 118735, Feb. 2020, doi: 10.1016/j.jclepro.2019.118735.

[41] T. Blu, P. Thevenaz, and M. Unser, “Linear Interpolation Revitalized,” *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 710–719, May 2004, doi: 10.1109/TIP.2004.826093.

[42] E. Y. Boateng, J. Otoo, and D. A. Abaye, “Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review,” *J. Data Anal. Inf. Process.*, vol. 08, no. 04, pp. 341–357, Sep. 2020, doi: 10.4236/jdaip.2020.84020.

[43] P. Cunningham and S. J. Delany, “k-Nearest Neighbour Classifiers - A Tutorial,” *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2022, doi: 10.1145/3459665.

[44] A. R. Lubis, M. Lubis, and A.- Khowarizmi, “Optimization of distance formula in K-Nearest Neighbor method,” *Bull. Electr. Eng. Informatics*, vol. 9, no. 1, pp. 326–338, Feb. 2020, doi: 10.11591/eei.v9i1.1464.

[45] L. M. Sinaga, Sawaluddin, and S. Suwilo, “Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, p. 012106, Jan. 2020, doi: 10.1088/1757-899X/725/1/012106.

[46] W. Li, Y. Chen, and Y. Song, “Boosted K-nearest neighbor classifiers based on fuzzy granules,” *Knowledge-Based Syst.*, vol. 195, p. 105606, May 2020, doi: 10.1016/j.knosys.2020.105606.

[47] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, “LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM,” *Knowledge-Based Syst.*, vol. 196, p. 105845, May 2020, doi: 10.1016/j.knosys.2020.105845.

[48] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, pp. 1255–1260, doi: 10.1109/ICCS45141.2019.9065747.

[49] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *J. Appl. Sci. Technol. Trends*, vol. 1, no. 4, pp. 140–147, 2020, doi: 10.38094/jastt1457.

[50] S. U. Mamatha *et al.*, “Multi-linear regression of triple diffusive convectively heated boundary layer flow with suction and injection: Lie group transformations,” *Int. J. Mod. Phys. B*, vol. 37, no. 01, Jan, p. 234, 2023, doi: 10.1142/S0217979223500078.

[51] F. Elmaz, Ö. Yücel, and A. Y. Mutlu, “Predictive modeling of biomass gasification with machine learning-based regression methods,” *Energy*, vol. 191, p. 116541, Jan. 2020, doi: 10.1016/j.energy.2019.116541.

[52] N. Shrestha, “Detecting Multicollinearity in Regression Analysis,” *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.

[53] M. Sholeh, E. K. Nurnawati, and U. Lestari, “Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 8, no. 1, pp. 10–21, Jan. 2023, doi: 10.14421/jiska.2023.8.1.10-21.

[54] A. Soy Temür and Ş. Yıldız, “Comparison of Forecasting Performance of ARIMA LSTM and HYBRID Models for The Sales Volume Budget of a Manufacturing Enterprise,” *Istanbul Bus. Res.*, vol. 50, no. 1, pp. 15–46, May 2021, doi: 10.26650/ibr.2021.51.0117.

[55] L. Wang, Y. Xia, and Y. Lu, “A Novel Forecasting Approach by the GA-SVR-GRNN Hybrid Deep Learning Algorithm for Oil Future Prices,” *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Aug. 2022, doi: 10.1155/2022/4952215.

[56] Z. Khan, T. Hussain, A. Ullah, S. Rho, M. Lee, and S. Baik, “Towards Efficient Electricity Forecasting in Residential and Commercial Buildings: A Novel Hybrid CNN with a LSTM-AE based Framework,” *Sensors*, vol. 20, no. 5, p. 1399, Mar. 2020, doi: 10.3390/s20051399.

[57] A. F. Adekoya, I. K. Nti, and B. A. Weyori, “Long Short-Term Memory Network for Predicting Exchange Rate of the Ghanaian Cedi,” *FinTech*, vol. 1, no. 1, pp. 25–43, Dec. 2021, doi: 10.3390/fintech1010002.

[58] D. Matzke and E.-J. Wagenmakers, “Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis,” *Psychon. Bull. Rev.*, vol. 16, no. 5, pp. 798–817, Oct. 2009, doi: 10.3758/PBR.16.5.798.

[59] M.-L. Zhang and Z.-H. Zhou, “ML-KNN: A lazy learning approach to multi-label learning,” *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007, doi: 10.1016/j.patcog.2006.12.019.

[60] D. Zheng, B. Qin, Y. Li, and A. Tian, “Cloud-Assisted Attribute-Based Data Sharing with Efficient User Revocation in the Internet of Things,” *IEEE Wirel. Commun.*, vol. 27, no. 3, pp. 18–23, Jun. 2020, doi: 10.1109/MWC.001.1900433.

[61] S. Mancini, V. I. Man’ko, and P. Tombesi, “Wigner function and probability distribution for shifted and squeezed quadratures,” *Quantum Semiclassical Opt. J. Eur. Opt. Soc. Part B*, vol. 7, no. 4, pp. 615–623, Aug. 1995, doi: 10.1088/1355-5111/7/4/016.

[62] B. C. Kelly, “Some Aspects of Measurement Error in Linear Regression of Astronomical Data,” *Astrophys. J.*, vol. 665, no. 2, pp. 1489–1506, Aug. 2007, doi: 10.1086/519947.

[63] N. Hofstra, M. Haylock, M. New, P. Jones, and C. Frei, "Comparison of six methods for the interpolation of daily, European climate data," *J. Geophys. Res. Atmos.*, vol. 113, no. D21, p. D21110, Nov. 2008, doi: 10.1029/2008JD010100.

[64] W. Sun and F.-J. Chang, "Empowering Greenhouse Cultivation: Dynamic Factors and Machine Learning Unite for Advanced Microclimate Prediction," *Water*, vol. 15, no. 20, p. 3548, Oct. 2023, doi: 10.3390/w15203548.