Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 1, April 2024, pp. 233~244
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i1.pp233-244

Improving ventilation classification in under-actuated zones: a k-nearest neighbor and data preprocessing approach

Yaddarabullah[1], Aedah Abd Rahman[2], Amna Saad[3]
[1]Department of Informatics, Universitas Trilogi, Jakarta, Indonesia
[2]School of Science and Technology, Asia e University, Subang Jaya, Malaysia
[3]Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Dec 12, 2023
Revised Jan 17, 2024
Accepted Jan 19, 2024

Keywords:

Cumulative moving average
Data preprocessing
Kalman filter
K-nearest neighbor
Occupant position
Under-actuated zone

ABSTRACT

This study investigates the use of k-nearest neighbors (k-NN) for classifying occupant positions in under-actuated zones, aiming to enhance ventilation control. The focus is on evaluating different data preprocessing techniques, particularly cumulative moving average (CMA), Kalman filtering (KF), and their combination, to boost the k-NN model's reliability and accuracy. The research uses received signal strength indicator (RSSI) data in a controlled setting. The methodology involves dividing the dataset into training and testing subsets and using root mean squared error (RMSE) to determine the best k value for model validation. The study performs a comparative analysis of the k-NN model's performance with both original and preprocessed RSSI data, focusing on metrics such as accuracy, precision, recall, F1-score, and RMSE. The findings emphasize the significant impact of the combined CMA-KF preprocessing technique in improving the model's accuracy and reliability. Specifically, this approach achieved an accuracy of 98.58%. The RMSE values are particularly noteworthy, exhibiting a perfect fit (RMSE of 0) for training data and a remarkably low RMSE of 0.119 for testing data, confirming the model's high accuracy and predictive capability.

Corresponding Author:

Yaddarabullah
Department of Informatics, Universitas Trilogi
Jakarta, Indonesia
Email: yaddarabullah@trilogi.ac.id

1. INTRODUCTION

Indoor air quality (IAQ) is a significant determinant of human health and wellbeing. Poor IAQ can lead to various health concerns, including respiratory problems, allergies, asthma, and respiratory diseases [1]. The primary cause of IAQ problems is indoor pollution sources that release gases or particles into the air. Inadequate ventilation can raise indoor pollution levels by not bringing in enough outside air to dilute pollutants [2]. Therefore, it is essential to understand and control common pollutants indoors to reduce the risk of indoor health concerns. One of the critical factors that affect IAQ is ventilation. The use of mechanical ventilation systems is a common method to distribute fresh air uniformly and efficiently to all occupants in a building [3], [4]. However, certain areas within buildings, known as under-actuated zones, may not have well-controlled air distribution systems, resulting in variations in cooling load and poor climate control, which may negatively affect IAQ and energy efficiency. The under-actuated zones in buildings are areas where the ventilation system is not capable of controlling the air exchange rate. These areas can have poor air quality, which can negatively affect the occupants' health. The issue with under-actuated zones is that the ventilation system may not be able to deliver fresh air to all occupied areas due to limited capacity or poor control [5].

Several studies have been conducted on the topic of IAQ and ventilation systems in both fully-actuated and under-actuated zones. These studies have explored various approaches to improving IAQ, including occupant position-based ventilation classification using machine learning algorithms and wireless sensor. Miao *et al.* [6] offers a real-time occupancy monitoring system based on machine learning methods to decrease building energy use. The authors classified occupancy data using the k-nearest neighbor (k-NN) technique. The paper describes the system's deployment and assessment in an office building utilizing a wireless sensor network. The findings shown that the system can detect occupancy reliably and minimize energy usage by managing heating, ventilation, and air conditioning (HVAC) systems based on occupancy data. The study emphasizes the potential of machine learning algorithms, such as k-NN, to optimize HVAC systems based on occupancy patterns in order to create energy-efficient buildings. Study by Jin *et al.* [7] proposes a wireless sensor network-based system for detecting occupancy and monitoring energy use in residential construction. The authors proposed a rule-based occupancy detection system that determines the occupancy state of each room in a building using data collected from a wireless sensor network. They also used occupancy data to assess a building's energy consumption patterns and created an energy management system to optimize energy use. This study examines the possibilities of wireless sensor networks for occupancy detection and energy management in residential buildings, which can lead to enhanced energy efficiency and cost savings.

The research of Simma *et al.* [8] discussing real-time occupancy estimation using Wi-Fi Networks to Optimize HVAC Operation, explained about the performance of energy on infrastructure in the US is quite high but has not been able to adjust occupants. Although the topic seems to overlap some with our research concept, we differentiate ourselves by focusing on measuring variations in ventilation performance, data management processes, and grouping classifications in ventilation systems based on occupant detection data. Anand *et al.* [9] study recommends an occupancy-based energy management solution for commercial building HVAC systems. This study employs clustering algorithms for occupancy categorization and decision trees for occupancy prediction. The authors utilized occupancy data to improve the functioning of a building's HVAC systems by altering the temperature and ventilation rates based on occupancy patterns. This study offers a simulation-based evaluation of the proposed system, demonstrating that the system can achieve considerable energy savings while preserving interior thermal comfort.

Drira and Smith [10] presented a machine-learning-based technique for detecting occupancy in buildings using structural vibrations. The authors employed k-NN and support vector machine (SVM) methods to classify occupancy data based on building structure vibrations. The findings show that the proposed system properly detects occupancy and surpasses the standard infrared-based occupancy detection system in terms of accuracy and reliability. This study emphasizes the potential of machine learning algorithms for occupancy identification in buildings utilizing nonintrusive sensors, such as structural vibrations, which can provide a cost-effective and reliable solution for occupancy monitoring. Zhang *et al.* [11] presents a machine learning-based technique for building occupancy prediction to optimize energy management. For occupancy prediction, the authors employed the k-NN and decision tree algorithms. This study describes the deployment and assessment of the proposed system in an office building using a wireless sensor network. This study emphasizes the potential of machine learning algorithms, such as k-NN and decision trees, for building occupancy prediction and energy management, which may lead to greater energy efficiency and cost savings.

Although the studies described above address the optimization of ventilation systems based on occupancy information, they do not directly address the categorization of ventilation strategies based on occupant position in under-actuated zones. Under-actuated zones are characterized by low ceiling levels and reduced volume, and they are controlled by systems with mechanical ventilation [12]-[14]. HVAC systems with mechanical ventilation are used to control the indoor air parameters in these zones, but they may not be optimized for energy efficiency or IAQ [15]. An efficient method to classify the position of occupants based on ventilation can help optimize the ventilation and IAQ in under-actuated zones by adjusting the HVAC system settings based on the occupancy status of each zone. This can lead to improved IAQ, reduced energy consumption, and increased occupant comfort and productivity. It is necessary to have an efficient method to classify the position of occupants based on ventilation in order to optimize ventilation and IAQ in under-actuated zone.

According to prior research, the k-NN approach is the most often used method for occupant position classification based wireless signal. This machine-learning technique categorizes data points based on their similarities to other data points. The parameter ‘k’ which represents the number of nearest neighbors considered, plays a crucial role in the performance of the k-NN algorithm [16]. Selecting an appropriate value for 'k' is essential to achieve optimal results. When ‘k’ is set to a small value, the k-NN algorithm becomes more sensitive to outliers, potentially leading to erroneous classifications [17], [18]. However, setting 'k' to an excessively large value may cause the neighborhood to include an uneven distribution of data points from different classes. This can impact the ability of the algorithm to make accurate classifications by introducing noise from unrelated data points. Based on distance computations, the k-NN approach finds the nearest neighbors. If the characteristics are not scaled, those with greater scales may dominate those with smaller scales [19]. Scaling ensures that every feature contributes equally to distance computations, eliminating bias and guaranteeing that characteristics are compared equitably [20]. In k-NN, feature scaling can lead to an improved speed and more accurate results. By bringing the features to a similar scale, the outliers and extreme values had less influence on the distance calculations. Moreover, the challenge of imbalanced data distribution is a notable concern in k-NN classification [21]. Imbalanced data occur when one class is significantly overrepresented compared to others, which is a common scenario in real-world applications such as fraud detection, rare disease diagnosis, and anomaly detection [22]. The presence of imbalanced data can adversely affect the k-NN classification process because the algorithm may exhibit biased behavior toward the majority class. Consequently, this can lead to suboptimal performance in accurately predicting instances from a minority class [23]. Hence, it is essential to address the issue of an imbalanced data distribution before initiating the classification process to ensure fair and accurate predictions.

In the context of classification based on wireless signal data, a key challenge is the variance in the received signal strength indicator (RSSI). This variance is closely tied to the multipath fading effect, which stems from signal reflections, diffractions, and interference, resulting in fluctuations in the signal strength at the receiver end [24]. The presence of multipath fading introduces signal variations that manifest as fluctuations in RSSI measurements received by wireless devices. Given that the RSSI is a pivotal metric for assessing signal strength, its variability can lead to a range of issues, including unreliable connectivity, decreased data rates, and even dropped calls or data sessions [25]. Prior research has explored the utilization of filter techniques to counteract the negative effects of multipath fading on classification [26]. These techniques aim to mitigate the influence of multipath fading and enhance the quality of the RSSI measurements for more accurate classification outcomes.

This study proposes a k-NN method to classify occupant position-based ventilation in under-actuated zones. Signal data from a wireless access point was utilized for the classification of occupant position-based ventilation. This study utilizes Shannon atrophy to identify imbalanced data and root mean square error (RMSE) to determine the number of neighbors (*k*). The investigation extends to data preprocessing techniques, specifically focusing on RSSI filter techniques. This study employs three distinct filter techniques: cumulative moving average (CMA), Kalman filter (KF), and a hybrid approach combining CMA with KF. These techniques are employed to counteract the effects of multipath fading, which lead to RSSI variance. By applying these filter techniques, this study aims to enhance the quality of the RSSI measurements and subsequently improve the accuracy of the k-NN classification. Min-max normalization was used to bring the features to a common scale, ensuring fair and unbiased comparisons in the distance calculations of the k-NN algorithm. Hyperparameter tuning is another critical aspect that the study engages with, seeking to optimize the performance of k-NN classification. This involves the selection of hyperparameters, such as the number of neighbors (k) and wsix assignments, aiming to fine-tune the algorithm to improve the performance. The core analysis involved a comparative examination of classification accuracy under various scenarios. This comparison encompasses both the dataset with original RSSI and dataset with RSSI processed using the filter techniques. This study assesses the impact of these techniques on the accuracy of k-NN classification, ultimately determining the most effective approach for accurately classifying occupant positions based on their ventilation requirements within under-actuated zones.

2. METHOD

The dataset employed in this research, integral to understanding under-actuated zones, was meticulously derived from RSSI data. This rich dataset, focusing on indoor location determination through Wi-Fi signal measurements, is publicly accessible at [27]. The dataset at its core is centered on Wi-Fi signal data collected from strategically positioned access points within a two-story building configuration, offering a diverse range of measurement scenarios. To prepare for in-depth analysis, the dataset underwent a comprehensive data cleansing process. During this initial step, six distinct areas representing ventilation on the building's first floor were carefully selected, as detailed in Table 1. This meticulous selection process was crucial to ensure the relevance and accuracy of the data for its intended analyses. Additionally, a significant update was introduced by incorporating a new variable termed 'distance.' The calculation of this variable was done with precision using the provided coordinates of the Wi-Fi access points, thereby introducing a vital spatial dimension to the data. Each area contains multiple timestamps, reflecting the frequency of RSSI measurements taken within each respective area.

Table 1. Zone area observed

Area	Coordinate (X, Y)	Distance	Total timestamps
Zone 1	22, 17	28	317
Zone 2	19, 17	26	289
Zone 3	18, 17	25	309
Zone 4	23, 35	42	400
Zone 5	23, 34	41	400
Zone 6	32, 33	46	400

The methodology employed in the research followed a well-structured and comprehensive approach. The process commenced with extensive data preparation, which involved refining the dataset using various filtering techniques. To enhance the dataset's suitability for advanced analysis, a significant augmentation process was applied. This included the addition of new columns containing RSSI filter values obtained through three distinct methodologies: CMA, KF, and the combination of CMA with KF (CMA-KF) as the proposed method. This stage provided a strong foundation for the subsequent phases of the study. Moving forward, the research conducted a thorough examination of data imbalances and feature scaling to ensure that the dataset was well-prepared for the upcoming classification tasks. During the classification stage, the primary focus centered on the utilization of the k-NN method. This phase played a crucial role as it aimed to identify the positions of occupants, with a specific emphasis on their ventilation requirements—an essential aspect for understanding indoor environmental quality. To enhance the effectiveness of the classification process, hyperparameter tuning was employed. This entailed fine-tuning various parameters of the k-NN algorithm to achieve optimal performance, a critical step in guaranteeing accurate and reliable classification results. The third and final stage of the study was dedicated to conducting a comprehensive assessment of the classification performance. This evaluation extended beyond the original dataset with RSSI values and also encompassed the dataset enriched with the RSSI filters.

2.1. Received signal strength indicator filter

The RSSI filtering is used to increase the accuracy and reliability of wireless communication systems, particularly in range-based localization [28]. RSSI is typically measured in decibels (dB) and indicates the power level of the signal as it arrives at the receiver. By reducing the short-term variations caused by interference, multipath propagation, and other variables, it provides a more steady and reliable indication of the true signal intensity. In this study, filtering techniques such as the CMA, KF, and CMA with KF was used.

2.1.1. Cumulative moving average

The CMA is a type of moving average filter that can be used to smooth out and minimize the noise in the RSSI signal [29]. The CMA filter is a quick and easy approach to compute the average of a series of RSSI values over time. The filter operates by summing the RSSI values over a specific time and dividing them by the number of values in the period. When the current RSSI values are added, the oldest value is subtracted from the total, and the new average is computed [28]. CMA has the advantage of significantly accounting for past data by accounting for the most recent data point. Initialize the filter with the first RSSI value, set cumulative_sum to this value, and set the count to one. Then, for each new RSSI value received, update the cumulative_sum by adding the new value, increment the count, calculate the new CMA by dividing the cumulative_sum by count, and use this CMA value as the smoothed RSSI value. These procedures are repeated for consecutive RSSI measurements to maintain a running average that smooths out variations and provides a more reliable depiction of the signal intensity over time.

2.1.2. Kalman filter

KF is a popular approach for removing noise from RSSI signals and improving wireless communication system accuracy [30]. The KF is a state estimator that uses noisy observations to estimate unobserved variables. This is a recursive method that considers the history of the measurements. The KF linear models, which means that the transition from one state to the next and the translation from state to measurement must be linear transformations [31]. The approach begins by initializing the initial state estimate x_0 and the initial error covariance P_0. Subsequently, a prediction phase forecasts the upcoming state estimate, x_k|k-1, based on the system dynamics using matrices F and B, while forecasting error covariance, P_k|k-1, encapsulating system uncertainty with regard to process noise via matrix Q. The update step, which is critical for improving accuracy, involves calculating the Kalman gain, K_k, which gauges the measurement reliability and adjusts the state estimate based on the difference between the observed measurement, z_k, and the predicted measurement, H * x_k|k-1, accounting for measurement noise using matrix R. The smoothed RSSI value emerges as the updated state estimate, x_k|k, as the KF iterates through these steps for subsequent RSSI inputs. The KF adeptly refines RSSI estimations using this approach, accommodating to dynamic wireless signal conditions while guaranteeing optimal accuracy and robustness in data processing.

2.1.3. Proposed method

The present study intends to employ a combined filtering technique that merges the advantages of CMA and KF to enhance the accuracy and dependability of RSSI measurements. The CMA is a swift and efficient method of averaging a series of RSSI values, enhancing the signal by diminishing short-term fluctuations resulting from interference or multipath propagation. It achieves this by continuously updating a cumulative sum with new RSSI values, taking into account the most recent data point. This running average effectively eliminates erratic variations, presenting a more stable depiction of the signal's strength. On the other hand, the KF is a more complex state estimator that refines the RSSI signal by considering the history of measurements and making predictions based on linear models. It begins with an initial estimate and error covariance, and then enters a cycle of prediction and updating. In each cycle, it adjusts the state estimate based on new observations while accounting for potential noise in the measurements, resulting in a refined and optimized RSSI reading. The process commences by applying the CMA to the RSSI dataset, followed by feeding the resulting CMA values into the KF. This hybrid approach seeks to leverage the promptness and straightforwardness of the CMA with the forecasting and optimizing capabilities of the KF, culminating in a doubly smoothed RSSI value.

2.2. Analyze imbalance data

The number of classes in the output variable is determined by the number of ventilations. In this study, a corner room with three ventilations is used. Analyzing unbalanced data assists in assessing the significance of attributes for various classes. When the data are skewed, models may overrely on traits that correlate with the majority class while ignoring those that are relevant for minority classes [22]. Shannon entropy is a measure of the uncertainty or information content of a random variable and can be used to analyze class imbalance in a dataset [32]. The entropy of random variable X is defined as follows:

$$H(X) = -\sum p(x) \,.\log p(x) \tag{1}$$

P(x) denotes the probability of the result x. In the context of datasets, particularly those used in machine learning and statistics, entropy is a crucial concept for understanding the balance and diversity of class distribution. When the entropy number is low, it implies that the dataset's class distribution is strongly skewed towards one or a few classes. This skewness is an indicator of imbalance in the dataset. For instance, in a binary classification problem, if a significant majority of data points belong to one class with only a small fraction representing the other, the dataset will exhibit low entropy. Such an imbalance can lead to biased models that perform well on the majority class but poorly on the minority class, as the model's exposure to the minority class is limited. Conversely, a high entropy number suggests a more uniform distribution of classes, indicating a balanced dataset where classes are more evenly represented. In this scenario, each class in the dataset has a similar probability of occurrence, leading to a higher degree of uncertainty or randomness, which is what entropy measure. High entropy in a dataset is generally desirable for machine learning purposes, as it means that the model is exposed to a more diverse set of examples from each class, enabling it to learn more generalized patterns.

2.3. Feature scaling

The variables in the data have different units; therefore, the data were transformed to have the same units. The second step was to transform the data using the feature-scaling method. Feature scaling is a technique used to standardize the independent features present in data in a fixed range. This was performed to ensure that all features contributed equally to the model and to prevent features with larger values from dominating the model. Min-maxing is a statistical technique used to rescale numerical values into the (0,1) range [33]. Xmax and Xmin are the maximum and lowest values of the feature, respectively. When the value of X is the highest number in the column, the numerator equals the denominator; hence, the value of X' is 1. If the value of X is between 0 and 1, then the value of X' is between 0 and 1. The min-max normalization procedure preserves the correlations between the original data values. The formula for min-max is as (2).

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

2.4. Classification using k-NN

k-NN classification is a well-known non-parametric classification approach that incorporates the simplest algorithm in problem solving and is capable of grouping items in a certain class based on similarities and differences [34]. k-NN classification utilizes the principle of locating the neighbors (k) and categorization based on similarity. The initial technique is to add additional points to the training data to be utilized as a categorization of test data. The method for determine neighbors (k) using the RMSE. The RMSE calculates the average difference between the anticipated and actual values of a statistical model [34]. This is the standard deviation of the residuals in mathematics. The gap between the regression line and data points is represented by residuals. The RMSE measures the variance of these residuals, indicating the degree to which the observed data cluster around the expected values. Low RMSE values imply that the model fits the data well and provides more accurate predictions. The data obtained in the previous process are carried out using split datasets for the sharing of training data and testing data. This study using 80% data calculation for training data and 20% for testing data. The ratio 80% for training and 20% for testing are a commonly used ratio in many fields of machine learning [36]. Ecludian distance was used to calculate the closest distance between the training and test data using [37].

2.5. Hyperparameter tuning

Hyperparameter tuning is an important optimization technique in machine learning algorithms, particularly for k-NN [38]. A grid search is essential for selecting the best hyperparameter configuration, which includes the value of 'k' and the wsix scheme used in k-NN [39]. The wsix scheme specifies the proportional wsix assigned to neighbor contributions, while the k value denotes the number of nearest neighbors evaluated during prediction. Performing hyperparameter tuning using grid search includes multiple steps: First, an exhaustive parameter grid is defined, with k values ranging and wsix scheme options (uniform and distance). Second, the dataset is partitioned into training and validation sets, allowing for the evaluation of model performance under various hyperparameters. The fundamental grid search procedure comprises iteratively training k-NN models on the training data using various k and wsix scheme combinations, then analyzing their performance on the validation set using relevant evaluation measures. This continual examination across the parameter grid provides information about the most effective hyperparameter arrangement. Choosing the best hyperparameters entails determining the combination that produces the best validation set performance. Finally, the hyperparameters chosen (k value and wsix scheme) are used to train a final k-NN model on the entire training dataset.

2.6. Performance measurement

Several k-NN classification performance measurements include the accuracy, precision, recall, F1 score, RMSE training, and RMSE testing [40]. The accuracy of the k-NN classifier was determined as the proportion of correctly classified instances divided by the total number of examples in the dataset. This provided a comprehensive overview of the performance of the model. To provide a detailed overview of the classifier performance, the confusion matrix displays the number of true positives, true negatives, false positives, and false negatives. It assists in examining the model's ability to effectively categorize instances and detect potential issues, such as misclassification of specific classes. Precision is defined as the proportion of correctly predicted positive cases (true positives) out of all projected positive instances (true positives + false positives), and reflects the model's ability to prevent false positives. The proportion of correctly predicted positive cases (true positives) out of the total number of actual positive instances (true positives + false negatives) was calculated using recall. It assesses the model's ability to identify all positive situations while avoiding false negatives. The F1 score is the harmonic mean of the accuracy and recall values, and provides a balanced assessment that considers both metrics.

3. RESULTS AND DISCUSSION

The dataset utilized in this study encompassed a total of 2,115 records, which were distributed across six distinct areas based on their respective distances. This rich collection of data was efficiently stored in a comma-separated value (CSV) file format, facilitating ease of manipulation and analysis. The study was composed of four experimental scenarios, each delving into different aspects of data processing and analysis. The initial phase of experimentation focused on the application of filtering techniques to the RSSI data, a crucial metric in wireless networks. This phase involved computing the CMA filter and the KF, as well as exploring their combined application. The aim was to smooth out the noise and enhance the reliability of the RSSI data, setting a strong foundation for more accurate analyses. Subsequently, the study's focus shifted towards analyzing the imbalanced nature of the dataset through Shannon Entropy. This approach quantified the extent of imbalance in data distribution across various classes, providing critical insights for further data preprocessing. The third phase involved feature scaling of the dataset via min-max normalization. This process adjusted the range of data feature values, ensuring that all features were on a uniform scale, a critical step for the effective application of machine learning algorithms. The fourth phase of the study entailed a comprehensive evaluation of the k-NN algorithm's performance. Hyperparameter tuning was applied to the k-NN algorithm to optimize its performance, enhancing its suitability for classification tasks in this context. This comparison aimed to determine the most effective data preprocessing technique for improving the accuracy and efficiency of k-NN in the context of wireless network signal strength. The last phase is comparative of measurement between original RSSI with CMA, KF, and CMA-KF.

3.1. Analyze imbalanced data

The dataset under analysis exhibited a notably skewed class distribution, characterized by six distinct classes that demarcate different levels of distance-based ventilation. This skewness, or non-uniformity, is particularly significant as it reflects the varying degrees of ventilation distances, each class representing a unique category in this spectrum. The disparities in class representation are vividly detailed in Figure 1, which serves as a crucial graphical illustration of the dataset's composition. This figure not only quantitatively outlines the number of instances in each class but also provides a visual representation of the disparity, making the imbalance immediately apparent and understandable even to those not deeply versed in statistical analysis. The figure effectively elucidates the relative proportions of each class within the dataset, offering an insightful view into the dataset's compositional heterogeneity.

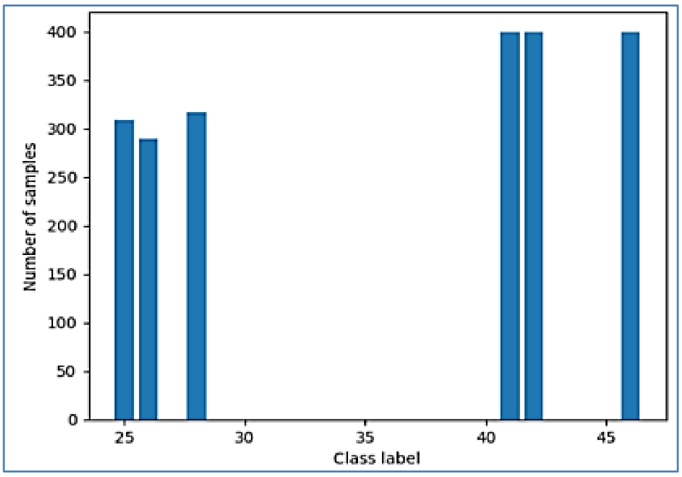

Figure 1. Distribution of classes in dataset

Following the visualization in Figure 1, Shannon Entropy was employed as a key metric to quantify the extent of imbalance within the dataset. Shannon Entropy, a concept derived from information theory, serves as a measure of uncertainty or randomness in the distribution of classes in a dataset. In this specific analysis, the calculated Shannon Entropy value was determined to be 0.9947. This value is substantial, pointing towards a lack of severe data imbalance in the dataset under consideration. Typically, entropy values close to 1 indicate a high level of uncertainty, which in the context of data distribution, translates to a more uniform spread of class instances. The significance of this high entropy value lies in its implication of a uniform distribution of classes within the dataset. It suggests that the dataset was characterized by a relatively equitable representation of instances across the various classes.

3.2. Performance of k-NN classification

Prior to selecting an optimal k-value for the k-NN algorithm, the sample data underwent a validation process using the root mean squared error (RMSE). The initial step involved dividing the sample data into two distinct sets: 80% for training and 20% for testing. The training set was utilized to construct the k-NN model, while the validation set was employed to assess the model's performance across various k-values. The primary objective of this experiment was to compare the performance of k-NN classification between datasets containing the original RSSI values and datasets enhanced with RSSI filters. The performance metrics, including accuracy, precision, recall, F1 score, error rate, and processing time, are presented in Table 2. The k-NN classification results using the original RSSI values yielded an accuracy of 90.54%. The model exhibited a precision of 0.91, indicating a relatively high proportion of true positive classifications among all predicted positive instances. Similarly, the recall score of 0.91 implied a strong capability to identify true positives from the actual positive instances. The F1-score of 0.90 reflected a balanced performance between precision and recall. However, it is noteworthy that the RMSE for testing was 0.307, signifying the average distance between predicted and actual distance values. These results suggest that the original RSSI values provide a reasonable foundation for position-based ventilation classification, although there is room for improvement, as indicated by the RMSE. The application of the CMA filter to the RSSI values led to a significant improvement in accuracy, reaching 97.87%. This enhancement was also evident in precision, recall, and F1-score, all of which increased to 0.98. Furthermore, the RMSE for testing decreased to 0.146, indicating reduced prediction errors. These improvements clearly demonstrate the efficacy of the CMA filter in reducing noise and enhancing the quality of RSSI data, ultimately resulting in more accurate occupant position classification. Similarly, when applying the KF to the RSSI values, an accuracy of 97.87% was achieved, with precision, recall, and F1-score all reaching 0.98, and the RMSE for testing remaining consistent at 0.146. These outcomes reinforce the positive impact of filtering techniques on classification performance. The KF's ability to handle dynamic systems and uncertainties appears to contribute significantly to the improved accuracy and reliability of the model. The proposed method, which combines both KF and CMA, yielded the highest accuracy among all techniques, reaching 98.58%. Precision, recall, and F1-score all achieved 0.99, underscoring the model's precision and reliability in classifying occupant positions. Moreover, the RMSE for testing dropped to 0.119, further substantiating the enhanced accuracy of the combined approach. The combination between KF and CMA leverages their respective strengths to provide superior classification results.

Table 2. Performance of k-NN classification

RSSI	*k* value	Accuracy	Precision	Recall	F1-score	RMSE training	RMSE testing
Original	2	90,54	0,91	0,91	0,90	0,175	0,307
CMA	2	97,87	0,98	0,98	0,98	0	0,146
KF	2	97,87	0,98	0,98	0,98	0	0,146
CMA and KF	2	98,58	0,99	0,99	0,99	0	0,119

According to Table 2, the results demonstrate that preprocessing RSSI data with filtering techniques significantly enhances the accuracy and reliability of occupant position classification for ventilation purposes in under-actuated zones. The CMA filter and KF independently improve the classification results by reducing noise and handling uncertainties, while their combination yields the best results. The consistently high precision, recall, and F1-score values across all filtering techniques underscore the models' ability to effectively distinguish between different occupant positions. Moreover, the decreasing RMSE values indicate that the predicted distances align closely with the actual distances, confirming the models' proficiency. The next experiment is comparing the performance of k-NN classification by using hyperparameter tuning.

Table 3 describes the performance result by applying hyperparameter tuning in k-NN classification. The hyperparameters investigated include the value of k (number of neighbors) and the weighting scheme (uniform or distance-based). The results indicate the selected hyperparameters for each preprocessing technique and provide insights into the impact of these selections on the classification performance. For the original RSSI data, the hyperparameter tuning resulted in selecting k=2 with a weighted scheme based on distance. The RMSE for testing was 0.243. This outcome suggests that considering the distances of the nearest neighbors weighted by their inverses (distance-based) contributed to better prediction performance compared to using a uniform weighting scheme. The low RMSE values for both training and testing indicate that the model fits well and generalizes effectively. For the RSSI data processed with CMA filter, the optimal hyperparameters were determined to be k=2 with a weighted scheme based on distance. The RMSE for testing was reduced to 0.166. The improved performance compared to the original RSSI data further supports the utility of filtering techniques in enhancing the quality of input data, resulting in more accurate predictions.

The selected hyperparameters again highlight the efficacy of distance-based weighting in improving prediction accuracy. The hyperparameter tuning for the RSSI data processed with a KF indicated that k=1 with a uniform weighting scheme was optimal. The RMSE for testing was 0.146. The result suggests that, in this case, considering the closest neighbor (k=1) with equal weighting across neighbors (uniform) led to the best prediction performance. The low RMSE value indicates that the model was able to accurately predict occupant positions, benefiting from the KF process. For the combined preprocessing of RSSI data with both CMA and KF, the hyperparameter tuning revealed that k=1 with a uniform weighting scheme was again the optimal choice. The RMSE for testing reached its lowest value at 0.119. This result underlines the significance of the cumulative effect of preprocessing techniques in yielding highly accurate predictions. The consistency in selecting k=1 with uniform weighting across the combined preprocessing methods suggests that the data quality enhancement outweighs the potential benefits of considering more neighbors or using distance-based weighting.

Table 3. Performance of k-NN classification using hyperparameter tuning

RSSI	Hyperparameter settings	Hyperparameter selected	RMSE training	RMSE testing
Original	k=1-50, weights=uniform, distance	k=2, weighted=distance	0	0,243
CMA	k=1-50, weights=uniform, distance	k=2, weighted=distance	0	0,166
KF	k=1-50, weights=uniform, distance	k=1, weighted=uniform	0	0,146
CMA and KF	k=1-50, weights=uniform, distance	k=1, weighted=uniform	0	0,119

The hyperparameter tuning results offer valuable insights into the optimal configurations for k-NN classification when applied to RSSI data with various preprocessing techniques. These configurations include the selection of two critical hyperparameters: the number of neighbors (k) and the weighting scheme (uniform or distance-based). These choices significantly influence the predictive performance of the k-NN model. One key observation across all preprocessing techniques is the preference for k=1, indicating that relying on the nearest neighbor's information plays a pivotal role in achieving accurate predictions. This suggests that in the context of indoor location determination and environmental quality assessment, considering the immediate proximity of data points is crucial for precise classification. Furthermore, the results underscore the substantial impact of data preprocessing techniques, specifically CMA and KF, on enhancing prediction accuracy. By applying these preprocessing methods to the RSSI data, the models exhibit improved predictive capabilities and greater reliability in their classifications. Comparing the RMSE values across different preprocessing techniques further highlights the positive influence of data enhancement on prediction accuracy. The consistently lower RMSE values for testing data indicate that the selected models generalize well and provide accurate predictions even when presented with unseen data.

3.3. Comparison of RSSI with filtering methods

Figure 2 presents a detailed analysis of RSSI measurements processed through three filtering methods (CMA, KF, and CMA-KF). These measurements are evaluated across six distinct zones, each representing different environmental conditions and signal strength dynamics. Starting with Zone 1, characterized by stable signal conditions, the CMA filter demonstrates its ability to create a steady graph line by smoothing minor fluctuations. In contrast, the KF method shows greater responsiveness to even slight changes, resulting in more pronounced fluctuations. The CMA-KF approach in this zone strikes a balance between smoothing and responsiveness, potentially providing the most accurate representation of the signal variations in this stable environment. Moving to Zone 2, where signal variability is higher, and the CMA-filtered data appears smoother but may lag behind immediate changes. Conversely, the KF-generated graph displays more pronounced fluctuations, adapting to rapidly changing signal conditions. The CMA-KF representation in this zone captures both the underlying trend and immediate signal changes effectively. Zone 3 represents an area prone to signal interference and obstructions, resulting in erratic RSSI readings. Here, the CMA filter produces a generalized, smoothed trend, potentially underestimating fluctuations. The KF-generated graph exhibits significant variability, reacting swiftly to changes. The CMA-KF approach balances noise reduction while indicating significant signal strength changes. Zone 4, a transitional area with predictable signal changes, shows the CMA filter slightly lagging behind, while the KF method closely follows the signal's trend. CMA-KF effectively smooths data while adapting to expected signal changes. In Zone 5, with weak signals and interference, CMA may oversmooth, while KF exhibits variability. CMA-KF provides a coherent representation by mitigating noise and capturing significant trends in signal degradation. Finally, Zone 6 presents dynamic conditions influenced by moving obstacles or varying interference sources. The CMA filter's excessive smoothing may miss rapid changes, while KF shows responsiveness. CMA-KF displays variability with a smoother curve, adapting to significant signal strength changes effectively.

The CMA method proves highly effective in mitigating the influence of transient fluctuations within RSSI values. Its primary advantage lies in its ability to attenuate the impact of rapid, fleeting anomalies in the data, a significant benefit when dealing with sudden and transient changes in RSSI readings. However, CMA's reliance on historical data introduces a certain level of latency, potentially hindering its responsiveness in swiftly changing signal environments. On the other hand, the KF operates dynamically, adapting to incoming data points while considering inherent uncertainties and variances in the signal data. This unique attribute enables the KF to provide precise estimations, especially in scenarios where the signal may be affected by factors such as physical obstructions, interference, or multi-path fading. KF's predictive nature allows it to respond promptly to genuine changes in signal strength, offering a significant advantage over CMA, which tends to exhibit delayed responses. The CMA-KF approach capitalizes on the strengths of both CMA and KF, presenting a balanced solution that mitigates the lag associated with CMA through the proactive adjustment capabilities of KF. This proposed hybrid method not only effectively smoothens out high-frequency noise but also adapts to real-time environmental changes. Consequently, the CMA-KF approach emerges as a nuanced solution that adeptly addresses the limitations of both CMA and KF, ultimately providing a more robust and responsive filter for processing RSSI data.

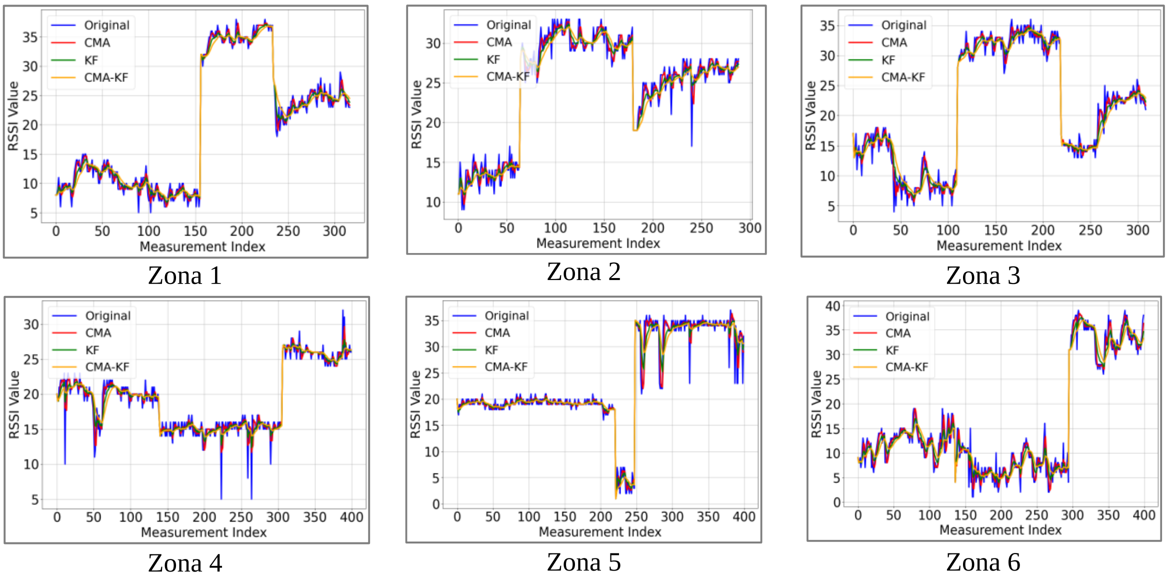

Figure 2. Measurement of RSSI with filter methods

4. CONCLUSION

In this study, we comprehensively explored the application of the k-NN classification algorithm for occupant position-based ventilation strategies in under-actuated zones, utilizing RSSI signals. Our research journey involved various phases, including employing preprocessing techniques and fine-tuning hyperparameters to significantly enhance classification accuracy and reliability. Notably, we achieved a substantial improvement in prediction accuracy through RSSI data preprocessing, specifically by applying CMA and KF. These filters effectively smoothed signal variability, significantly enhancing data consistency. CMA improved accuracy from the original 90.54% to an impressive 97.87%, while KF achieved a similar accuracy of 97.87%. The most remarkable result, however, emerged from our combined use of these filtering techniques – the CMA-KF approach, which outperformed all others with an accuracy of 98.58%. This underscores the vital role of data quality enhancement, clearly demonstrating CMA-KF's superiority. Hyperparameter tuning played a critical role in optimizing the k-NN model, involving adjustments to parameters like the number of neighbors (k) and the choice of weighting scheme (uniform or distance-based). This tailored tuning was essential due to the varying optimal configurations across different preprocessing methods. The practical implications are extensive, particularly in implementing occupant position-based ventilation strategies in under-actuated zones. Leveraging machine learning and historical RSSI data allows our predictive system to accurately anticipate room occupancy changes, enabling proactive HVAC system control for enhanced energy efficiency and indoor environmental quality. Future research should focus on integrating advanced techniques like convolutional neural networks (CNNs) to further improve the system's capability. CNNs can extract intricate patterns from RSSI sensor data, enhancing occupancy predictions.

ACKNOWLEDGEMENTS

The authors offer their sincerest appreciation to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for their generous financial support, which has been crucial in facilitating this study. The authors extend their deepest gratitude to the research assistants at Universitas Trilogi. The authors are also grateful to Universitas Trilogi for granting them access to the laboratories.

REFERENCES

[1] V. V. Tran, D. Park, and Y. C. Lee, "Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality," *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, pp. 1–27, 2020, doi: 10.3390/ijerph17082927.

[2] S. Sadrizadeh *et al.*, "Indoor air quality and health in schools: A critical review for developing the roadmap for the future school environment," *Journal of Building Engineering*, vol. 57, 2022, doi: 10.1016/j.jobe.2022.104908.

[3] H. B. Awbi, "Ventilation for good indoor air quality and energy efficiency," *Energy Procedia*, vol. 112, no. October 2016, pp. 277–286, 2017, doi: 10.1016/j.egypro.2017.03.1098.

[4] X. Sui, Z. Tian, H. Liu, H. Chen, and D. Wang, "Field measurements on indoor air quality of a residential building in Xi'an under different ventilation modes in winter," *Journal of Building Engineering*, vol. 42, p. 103040, 2021, doi: 10.1016/j.jobe.2021.103040.

[5] J. Brooks, S. Kumar, S. Goyal, R. Subramany, and P. Barooah, "Energy-efficient control of under-actuated HVAC zones in commercial buildings," *EnergyBuild*, vol. 93, pp. 160–168, 2015, doi: 10.1016/j.enbuild.2015.01.050.

[6] Y. Miao, A. Hunter, and I. Georgilas, "An occupancy mapping method based on K-nearest neighbours," *Sensors*, vol. 22, no. 1, 2022, doi: 10.3390/s22010139.

[7] X. Jin, G. Wang, Y. Song, and C. Sun, "Smart building energy management based on network occupancy sensing," *Journal of International Council on Electrical Engineering*, vol. 8, no. 1, pp. 30–36, 2018, doi: 10.1080/22348972.2018.1462608.

[8] K. C. J. Simma, A. Mammoli, and S. M. Bogus, "Real-time occupancy estimation using wifi network to optimize HVAC operation," *Procedia Computer Science*, vol. 155, pp. 495–502, 2019, doi: 10.1016/j.procs.2019.08.069.

[9] P. Anand, D. Cheong, and C. Sekhar, "A review of occupancy-based building energy and IEQ controls and its future post-COVID," *Science of the Total Environment*, vol. 804, p. 150249, Jan. 2022, doi: 10.1016/j.scitotenv.2021.150249.

[10] S. Drira and I. F. C. Smith, "A framework for occupancy detection and tracking using floor-vibration signals," *Mechanical Systems and Signal Processing*, vol. 168, p. 108472, 2022, doi: 10.1016/j.ymssp.2021.108472.

[11] W. Zhang, Y. Wu, and J. K. Calautit, "A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment," *Renewable and Sustainable Energy Reviews*, vol. 167, p. 112704, 2022, doi: 10.1016/j.rser.2022.112704.

[12] G. Zucker, A. Sporr, A. Garrido-Marijuan, T. Ferhatbegovic, and R. Hofmann, "A ventilation system controller based on pressure-drop and CO2 models," *The Energy Building*, vol. 155, pp. 378–389, 2017, doi: 10.1016/j.enbuild.2017.09.041.

[13] P. Anand, D. Cheong, and C. Sekhar, "Computation of zone-level ventilation requirement based on actual occupancy, plug and lighting load information," *Indoor and Built Environment*, vol. 29, no. 4, pp. 558–574, 2020, doi: 10.1177/1420326X19875802.

[14] Y. Shi, X. Li, and S. Sadatiseyedmahalleh, "Influence of building envelope type on the minimum mechanical ventilation rate to achieve a positive indoor air pressure," *E3S Web Conferences*, vol. 172, p. 5002, 2020, doi: 10.1051/e3sconf/202017205002.

[15] S. Dhanalakshmi, M. Poongothai, and K. Sharma, "IoT based indoor air quality and smart energy management for HVAC system," *Procedia Computer Science*, vol. 171, pp. 1800–1809, 2020, doi: 10.1016/j.procs.2020.04.193.

[16] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," *International Conference on Intelligent Computing and Control Systems*, 2019, pp. 1255–1260, doi: 10.1109/ICCS45141.2019.9065747.

[17] H. Xu, L. Zhang, P. Li, and F. Zhu, "Outlier detection algorithm based on k-nearest neighbors-local outlier factor," *Journal of Algorithms and Computational Technology*, vol. 16, 2022, doi: 10.1177/17483026221078111.

[18] P. O. Olukanmi and B. Twala, "K-means-sharp: Modified centroid update for outlier-robust k-means clustering," in *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, 2017, pp. 14–19, doi: 10.1109/RoboMech.2017.8261116.

[19] S. Basak and M. Huber, "Evolutionary feature scaling in k-nearest neighbors based on label dispersion minimization," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 928–935, doi: 10.1109/SMC42975.2020.9282834.

[20] S. A. Zadeh, M. Ghadiri, V. Mirrokni, and M. Zadimoghaddam, "Scalable feature selection via distributed diversity maximization," *31st AAAI Conference on Artificial Intelligence AAAI 2017*, pp. 2876–2883, 2017, doi: 10.1609/aaai.v31i1.10926.

[21] P. Nair and I. Kashyap, "Hybrid pre-processing technique for handling imbalanced data and detecting outliers for KNN classifier," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 460–464, doi: 10.1109/COMITCon.2019.8862250.

[22] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Jornal Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0192-5.

[23] Z. Shi, "Improving k-nearest neighbors algorithm for imbalanced data classification," *IOP IOP Conference Series: Materials Science and Engineering*, vol. 719, no. 1, 2020, doi: 10.1088/1757-899X/719/1/012072.

[24] D. A. Rafina, P. Kristalina, and A. Sudarsono, "Modified iterated extended Kalman filter for mobile cooperative tracking system," *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, vol. 7, no. 3, pp. 980–992, 2017, doi: 10.18517/ijaseit.7.3.2657.

[25] J. A. Santana, E. Macías, Á. Suárez, D. Marrero, and V. Mena, "Adaptive estimation of WiFi RSSI and its impact over advanced wireless services," *Mobile Networks and Applications*, vol. 22, no. 6, pp. 1100–1112, 2017, doi: 10.1007/s11036-016-0729-1.

[26] M. M. Agel, M. H. Habaebi, and M. R. Islam, "Mitigation of multipath fading in indoor radiometric fingerprinting systems," *Computers and Electrical Engineering*, vol. 73, pp. 46–57, 2019, doi: 10.1016/j.compeleceng.2018.11.002.

[27] *Indoor location determination with RSSI*, www.kaggle.com [Online]. Available: https://www.kaggle.com/datasets/amirma/indoor-location-determination-with-rssi

[28] M. A. Koledoye, D. D. Martini, S. Rigoni, and T. Facchinetti, "A comparison of RSSI filtering techniques for range-based localization," in *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2018, pp. 761–767, doi: 10.1109/ETFA.2018.8502556.

[29] D. T. Thinh, N. B. H. Quan, and N. Maneetien, "Implementation of moving average filter on STM32F4 for vibration sensor application," in *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, 2018, pp. 627–631, doi: 10.1109/GTSD.2018.8595630.

[30] P. H. Abreu, J. Xavier, D. C. Silva, L. P. Reis, and M. Petry, "Using Kalman filters to reduce noise from RFID location system," *The Scientific World Journal*, vol. 2014, p. 796279, 2014, doi: 10.1155/2014/796279.

[31] D. Peng, C. Wen, and M. Lv, "Design of a high-order Kalman filter for state and measurement of a class of nonlinear systems based on kronecker product augmented dimension," *Sensors*, vol. 23, no. 6, 2023, doi: 10.3390/s23062894.

[32] A. Bulinski and D. Dimitrov, "Statistical estimation of the shannon entropy," *Acta Mathematica Sinica, English Series*, vol. 35, no. 1, pp. 17–46, 2019, doi: 10.1007/s10114-018-7440-z.

[33] S. Sinsomboonthong, “Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification,” *International Journal of Mathematics and Mathematical*, p. 3584406, 2022, doi: 10.1155/2022/3584406.
[34] M. Novitasari, Yaddarabullah, S. D. H. Permana, and E. D. Krishnasari, “Classification of house buildings based on land size using the K-nearest neighbor algorithm,” *AIP Conference Proceedings*, vol. 2499, no. 1, p. 50010, Nov. 2022, doi: 10.1063/5.0104960.
[35] K. Tyagi, C. Rane, Harshvardhan, and M. Manry, “Chapter 4 - regression analysis,” *Academic Press*, 2022, pp. 53–63, doi: 10.1016/B978-0-12-824054-0.00007-1.
[36] J. Grus, "*Data Science from Scratch,"* 2nd Edition, O’Reilly Media, Inc., 2019, [Online]. Available: https://www.oreilly.com/library/view/data-science-from/9781492041122/
[37] M. Nishom, S. F. Handayani, and D. Dairoh, “Pillar Algorithm in K-means method for identification health human resources availability profile in Central Java,” *JUITA : Jurnal Informatika*, vol. 9, no. 2, p. 145, 2021, doi: 10.30595/juita.v9i2.9860.
[38] E. Jääsaari, V. Hyvönen, and T. Roos, “Efficient autotuning of hyperparameters in approximate nearest neighbor search BT - advances in knowledge discovery and data mining,” *Springer International Publishing*, 2019, pp. 590–602.
[39] R. Ghawi and J. Pfeffer, “Efficient hyperparameter tuning with grid search for text categorization using k-NN approach with BM25 similarity,” *Open Computer Science*, vol. 9, no. 1, pp. 160–180, 2019, doi: doi:10.1515/comp-2019-0011.
[40] J. Meghana, J. Hanumanthappa, and S. P. S. Prakash, “Performance comparison of machine learning algorithms for data aggregation in social internet of things,” *Global Transitions Proceedings*, vol. 2, no. 2, pp. 212–219, 2021, doi: 10.1016/j.gltp.2021.08.032.

BIOGRAPHIES OF AUTHORS

Yaddarabullah is lecturer at Informatics Department, Universitas Trilogi, Jakarta, Indonesia. He holds a master degree in Computer Engineering with specialization in applied computing. In 2021, he obtained his Ph.D. in Information Communication Technology from Asia e-University, Malaysia, with a specific focus on artificial neural networks in building management systems. His research specialization spans a broad spectrum of cutting-edge technologies, including machine learning, distributed computing, mobile computing, the internet of things, and network engineering. He has published 52 papers spanning prestigious journals along with conference proceedings. His published articles have received 180 citations in Google Scholar with 8 h-index, and 32 citations in Scopus with a 3 h-index. He can be contacted at email: yaddarabullah@trilogi.ac.id.

Aedah Abd Rahman is Associate Professor at School of Science and Technology, Asia e-University (AeU), Selangor, Malaysia. She has supervised and co-supervised more than 30 Ph.D. students. She received a Ph.D. in Computer Science from Universiti Teknologi Malaysia. She is Head of Department ICTS & Asian Centre of e Learning (ACE), and Head of Unit Accreditation of Prior Experential Learning (APEL) in AeU. She has certified trainer in Human Resource Development Corporation and senior member IEEE. Her research interests include software engineering, software quality, and project management. Her published 70 articles have received 328 citations in Google Scholar with 10 h-index, and 37 citations in Scopus with 5 h-index. She can be contacted at email: aedah.abdrahman@aeu.edu.my.

Amna Saad is Senior Lecturer in the System and Networking Department at Universiti Kuala Lumpur Malaysian Institute of Information Technology (MIIT), Kuala Lumpur, Malaysia. She received a Ph.D. in Computer Science from the Loughborough University and worked as an Application Development Manager, Value Added Network Services, Enterprise Network Solutions, COINS, Telekom Malaysia (TM) Berhad, before she joined Universiti Kuala Lumpur in 2003. Her research interests include network performance analysis including VoIP performance, big data and data analytic, IoT and sensory devices. Her published 70 articles have received 74 citations in Google Scholar with 6 h-index, and 27 citations in Scopus with 3 h-index. She can be contacted at email: amna@unikl.edu.my.