Al-Ishlah: Jurnal Pendidikan
Vol.18, 1 (March, 2026), pp. 1076-1091
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v18i1.9624

Development of Smart Assessment System for Evaluating Maritime English Competence Using Machine Learning

Aprizawati[1], Romadhoni[2], Budhisantoso[3]

[1] Politeknik Negeri Bengkalis, Bengkalis, Indonesia; aprizawati@polbeng.ac.id
[2] Politeknik Negeri Bengkalis, Bengkalis, Indonesia; romadhoni@polbeng.ac.id
[3] Politeknik Negeri Bengkalis, Bengkalis, Indonesia; budhisantoso@polbeng.ac.id

ARTICLE INFO

Keywords:

competency evaluation;
intelligent assessment systems;
machine learning;
maritime education;
maritime English

Article history:

Received 2021-08-14
Revised 2021-11-12
Accepted 2022-01-17

ABSTRACT

Maritime English proficiency assessment is essential for cadets and maritime professionals, yet manual scoring can be time-consuming and prone to inter-rater variability. This study proposes a web-based smart assessment system that integrates machine learning to classify Maritime English proficiency into Beginner/Intermediate/Advanced using four feature scores: listening, reading, writing, and speaking. The dataset used in this work is simulated (1,000 records) for proof-of-concept evaluation because access to large, standardized real examination data was limited and required institutional clearance; simulation enables controlled class balance and repeatable experimentation. Class labels are generated using rubric-based threshold rules, and the labelling scheme is validated by two Maritime English examiners who review the thresholds and independently rate a random subset of 200 samples; agreement is quantified using Cohen's kappa (κ) to ensure reliability. We adopt an 80/20 hold-out split and apply stratified 5-fold cross-validation on the training set for model selection, using grid-search hyperparameter tuning. We compare Support Vector Machine (SVM) and Random Forest and report accuracy, precision, recall, macro-F1, and brief per-class performance for Beginner/Intermediate/Advanced. SVM achieves 92% accuracy with macro-F1 = 0.905, outperforming Random Forest (89%, macro-F1 = 0.875). Future work will validate the system using real assessment datasets in operational training settings.

Corresponding Author:
Aprizawati
Politeknik Negeri Bengkalis, Bengkalis, Indonesia; aprizawati@polbeng.ac.id

1. INTRODUCTION

English is the working language of international shipping, and safety-critical communication at sea requires standardized phraseology. The International Maritime Organization (IMO) introduced the Standard Marine Communication Phrases (SMCP) to support clear and unambiguous operational communication, and Maritime English training, therefore, becomes a key component in maritime education and competence development (Frolova, 2020; Kulikova, 2024). In Indonesian maritime vocational education, Maritime English is commonly assessed through written examinations and oral

interviews; however, such human-centered assessment can be time-consuming, costly to administer at scale, and susceptible to subjectivity when different examiners apply scoring criteria with varying strictness (Enrico et al., 2011). These constraints create a practical need for assessment workflows that are faster, more consistent, and easier to audit while still reflecting performance across the core language skills required in maritime operations.

Technology-assisted assessment (e.g., online testing and simulation-based learning environments) has been increasingly adopted in maritime education, enabling more efficient administration and digital record-keeping (Salman, 2021; Tutie, 2023; Iie et al., 2025). Nevertheless, many implementations remain "digital substitutes" of conventional testing, collecting scores without providing systematic, data-driven interpretation that supports reliable proficiency decisions and targeted feedback. Recent studies on automated assessment and AI-supported feedback highlight the potential of intelligent systems to improve efficiency and reduce subjective errors in evaluation, but they also emphasize the need for careful design, validation, and transparent evaluation protocols in educational assessment (Michał, 2025). In the specific domain of Maritime English, where communication demands are standardized and safety-critical, there is still limited work that integrates an end-to-end assessment workflow with an objective proficiency classification engine that instructors can use for rapid screening and remediation decisions.

Machine learning (ML) offers a practical approach to learning patterns from multi-skill performance indicators and producing consistent classification outcomes. Prior work has shown that classical ML models such as Support Vector Machine (SVM) and Random Forest can achieve strong performance in classification tasks and can be applied to educational and language-related settings when features are well-defined, and evaluation is rigorous (Marcin et al., 2021; Samer et al., 2024). Motivated by this opportunity and the assessment limitations described above, this study focuses on an intelligent assessment approach for Maritime English based on four skill scores: listening, reading, writing, and speaking, within a web workflow that supports timely feedback for instructors and learners (Hadeel & Mohammed, 2020; Fan, 2023).

The contribution of this paper is twofold. First, we present a system engineering contribution: a web-based smart assessment workflow that manages test sessions, captures four-skill scores, and delivers automated results via an integrated ML module. Second, we provide an ML evaluation contribution by comparing SVM and Random Forest for three-level proficiency classification (Beginner/Intermediate/Advanced) using a controlled experimental setup and standard multi-class metrics.

The objective of this study is to develop and evaluate a web-based intelligent assessment system that classifies Maritime English proficiency levels from four-skill test results with reliable and consistent outputs. The research questions are:

1. RQ1: Can ML models classify Maritime English proficiency levels (Beginner/Intermediate/Advanced) from four-skill scores with acceptable performance?
2. RQ2: How consistent are the model outputs with expert judgment when the same rubric and level definitions are applied?
3. RQ3: Is the proposed system feasible within a web workflow in terms of practical deployment (e.g., response time and operational usability)?

Given the need for efficient, consistent, and objective assessment in Maritime English, this study proposes an intelligent, web-based approach that integrates four-skill evaluation with machine learning–based proficiency classification. By leveraging structured performance data and established classification models, the proposed system seeks to enhance the reliability of assessment outcomes while supporting timely instructional decisions. In doing so, it aims to contribute to the advancement of technology-assisted language assessment in maritime education, particularly in contexts where standardized, safety-critical communication is essential.

2. METHODS

2.1 Design Science Research (DSR) Methodology

This study adopts a Design Science Research (DSR) approach because the main output is an IT artifact in the form of a web-based smart assessment system, accompanied by an empirical evaluation of its classification component. DSR is appropriate for studies that (i) identify a practical problem, (ii) design and build a solution artifact, and (iii) evaluate the artifact through systematic testing and performance evidence.

The research procedure follows six DSR stages:

1. Problem identification and motivation: Maritime English assessment requires a workflow that is efficient and consistent, while manual scoring can be time-consuming and potentially variable across raters.
2. Define objectives of the solution: develop an integrated web-based assessment workflow and an ML-based classifier that outputs proficiency levels (Beginner/Intermediate/Advanced) from four-skill scores.
3. Design and development: design the system architecture, database, user roles, and ML module; implement the web application and integrate the trained model for prediction and reporting.
4. Demonstration: deploy the prototype in a web workflow and demonstrate the end-to-end process (input scores → prediction → dashboard/report).
5. Evaluation: conduct evaluation in two separate strands (system evaluation and model evaluation) as described in Section 2.1.1 and Section 2.1.2.
6. Communication: report the artifact design, experimental protocol, and findings in this paper.

2.1.1 Strand A: System Development Evaluation

System evaluation assesses whether the developed artefact functions correctly and is feasible for operational use. The evaluation includes:

a. functional testing of main modules (authentication/roles, input forms, scoring storage, prediction output, dashboard/report generation),
b. performance testing (e.g., response time for prediction and report generation under typical use),
c. Usability testing (if conducted) describes participants, tasks, and instruments (e.g., SUS score or structured feedback). The evaluation outputs are test results, observed issues, and improvements applied to the prototype.

2.1.2 Strand B: Model Evaluation (Machine Learning)

Model evaluation focuses on the classification performance of ML models using the defined feature set (four-skill scores) and the labeling scheme (Beginner/Intermediate/Advanced). This strand specifies:

a. Data split and validation protocol (e.g., hold-out + stratified k-fold cross-validation on training set),
b. Hyperparameter tuning procedure (e.g., grid search with cross-validation), and
c. Reporting metrics for multi-class classification (accuracy, precision, recall, macro-F1, and per-class performance). The outputs of this strand are comparative results between models and evidence of consistency of performance across folds.

2.2 Research and Development Process for System Implementation

This research employs a Research and Development (R&D) approach to design and implement an intelligent web-based assessment system for evaluating Maritime English proficiency. The development process follows five main stages: (1) needs analysis, which identifies the functional and technical requirements of the system; (2) system design, focusing on the architectural structure, database schema, and user interface; (3) machine learning implementation, where algorithms such as Support Vector Machine (SVM) and Random Forest are applied to process and classify learners'

language data; (4) system testing, conducted to evaluate functionality, usability, and model performance; and (5) results evaluation, which evaluates the efficacy and accuracy of the intelligent assessment system. This methodical approach to research and development guarantees that the system is both technically sound and contextually relevant from a pedagogical standpoint.

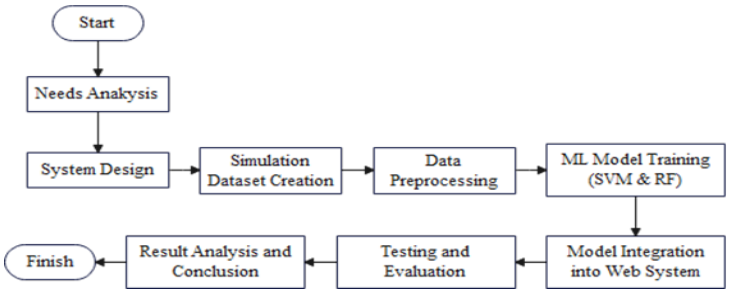

Figure 1. Research methodology flowchart design

2.3 System Architecture

This intelligent and adaptive web-based maritime English assessment is designed with the needs and characteristics of Nautical Cadets in mind. With this system, it is hoped that they can measure their learning progress more effectively and efficiently. Additionally, this system can also provide material recommendations tailored to the understanding level and needs of each nautical cadet. Thus, the implementation of machine learning technology in education can make a significant contribution to improving the quality of maritime English learning for nautical cadets (Sharma, 2023) . The system is web-based and has three main components:

a) Front-End Interface
Providing a user interface for test administration consists of listening, reading, writing, and speaking. With this system, it is hoped that naval cadets can learn maritime English more effectively and efficiently. With tailored material recommendations, their level of understanding is also expected to increase. Thus, the quality of maritime English learning for Naval Cadets can be continuously improved through the implementation of machine learning technology in education.

b) Back-End Server
Managing test data, processing results with ML models, and providing useful information for evaluating the progress of Naval Academy cadets' learning. Additionally, the back-end server can also provide feedback and recommendations to instructors regarding areas that need improvement in maritime English language learning. With a system that can automatically analyze the learning needs of naval cadets, it is hoped that their learning outcomes can be more optimal (Bolbot, 2022). Thus, the implementation of machine learning technology in education can make a positive contribution to improving the quality of maritime English learning for nautical cadets.

c) Machine Learning Engine
A module that classifies competency levels based on exam results. This machine learning engine can help identify the weaknesses and learning needs of nautical cadets (J. & n, 2023) . With the help of the Machine Learning Engine, instructors can more effectively design learning programs tailored to the individual needs of each nautical cadet. Thus, they can focus on areas that need improvement and enhance their maritime English skills more efficiently. With the implementation of this technology, it is hoped that there will be a significant improvement in the quality of maritime English education for Nautical Cadets.

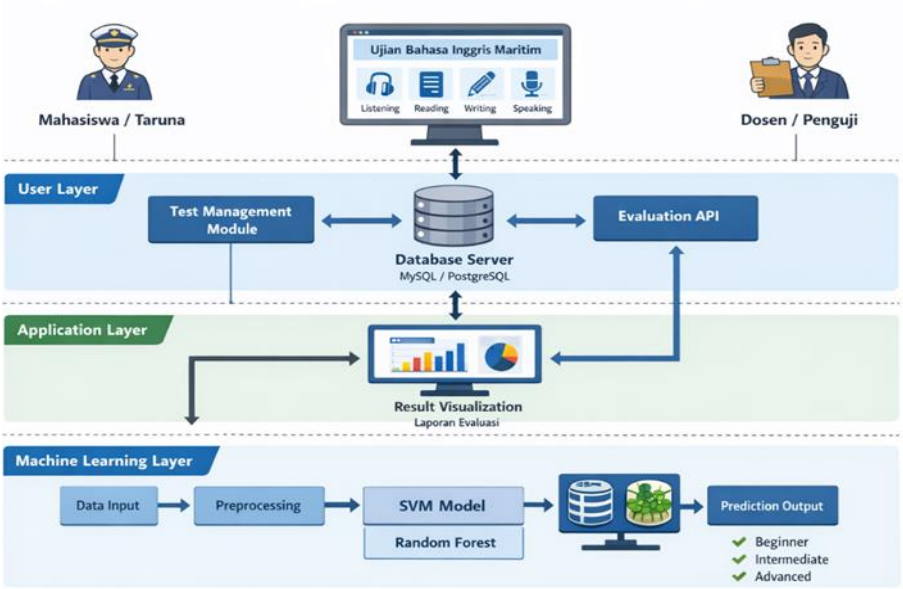

Figure 2. Intelligent Assessment System Architecture

2.4 Dataset and Preprocessing

The dataset used in this study is a simulated Maritime English assessment dataset consisting of 1,000 records, where each record contains four numeric skill scores: listening (L), reading (R), writing (W), and speaking (S). The pre-processing steps include data normalization and outlier removal. (Vaishali & Rupa, 2011). This process is carried out to ensure that the data used in training and testing is valid and accurate. After the pre-processing stage is complete, the model is trained using machine learning algorithms to improve the maritime English test results of Nautical cadets. (Yisi, 2025).

The simulation was employed for proof-of-concept evaluation because access to large, standardized cadet assessment data is restricted due to institutional clearance and privacy considerations; therefore, simulation enables controlled class balance and repeatable experimentation while the web-based assessment workflow is being developed. Data pre-processing is a crucial step in the machine learning model development process. (Ervin et al., 2022). Once the data is ready, the next step is to choose a machine learning algorithm that is suitable for the characteristics of the data and the goals you want to achieve. (Yong & algorithm, 2020).

2.4.1 Simulated data generation

To improve transparency, the simulated dataset was generated with the following assumptions. First, we sampled the intended proficiency group distribution to reflect three levels: Beginner (23.7%), Intermediate (33.0%), and Advanced (43.3%), matching the system-level distribution reported in the results section.

For each group, four-skill scores were generated from a correlated score profile to mimic the common dependency among skills in language assessment. Specifically, a latent proficiency factor was used to induce positive correlations among (L, R, W, S), and then small independent noise was added to represent measurement variability. Scores were generated on a 0–100 scale, clipped to remain within valid bounds. Group-wise score profiles were set so that Beginner, Intermediate, and Advanced have increasing central tendencies, while allowing moderate overlap across groups to represent realistic variation (e.g., a learner may be stronger in reading than speaking). This procedure produces four-skill score vectors that are internally consistent and suitable for evaluating the ML pipeline and the end-to-end system workflow.

This explicit thresholding clarifies that the labels are rule-derived. Consequently, the ML task in this study may partially learn the threshold function rather than latent language competence; this

limitation is acknowledged, and future work will validate the approach using real cadet assessment data and richer features (e.g., rubric sub-scores or performance artifacts for speaking/writing) to better capture authentic proficiency.

Train–test split and validation protocol

The dataset was split into 80% training and 20% testing.

To ensure stable model selection, we applied stratified 5-fold cross-validation on the training set during hyperparameter tuning, and then reported final performance on the held-out test set. This protocol prevents information leakage and provides a more reliable estimate of generalization than a single split.

2.4.2 Preprocessing

Two preprocessing steps were applied prior to model training. First, outliers were removed using the interquartile range (IQR) rule to reduce the influence of extreme values. Second, all four features were normalized using Min–Max normalization to map scores into [0.1][0.1][0.1], improving comparability across features and supporting stable learning. The normalization parameters were computed from the training data and applied to the test data to avoid data leakage, plan for validation with real data. Although simulation supports reproducible proof-of-concept evaluation, it is a high-risk limitation for reputable journals. Therefore, future work will conduct validation using real cadet assessment datasets (subject to institutional approval), including examiner-rated speaking/writing artifacts and formal inter-rater reliability reporting.

Labeling rule because the study uses score-based labels, the proficiency label yyy is defined explicitly from the average of the four skills:

$$AvgScore = \frac{L+R+W+S}{4} \quad (1)$$

Proficiency labels are then assigned using rubric-inspired thresholds:

a) Beginner (B) if AvgScore
b) Intermediate (I) if 60 ≤ AvgScore < 75
c) Advanced (A) if AvgScore ≥ 75

2.5 Ground Truth and Label Validation

Because the proficiency labels in this study are derived from explicit score-threshold rules, they should be treated as rule-based reference labels rather than absolute ground truth of language competence. To validate the labelling scheme and avoid contradictory claims, we conducted an examiner-based validation protocol focused on level assignment consistency. Two Maritime English examiners (E1–E2), each with experience in maritime vocational English assessment, reviewed the rubric thresholds and then independently assigned proficiency levels.
(Beginner/Intermediate/Advanced) to a random subset of 200 records. To reduce bias, examiners were blinded to the threshold-generated labels and model predictions; they received only the four-skill score profiles (listening, reading, writing, speaking) and the level definitions. The examiner labels were compared against the rule-based labels and between examiners.

Agreement was reported using (i) observed agreement (percentage of identical labels) and (ii) chance-corrected inter-rater reliability using weighted Cohen's kappa (ordinal weighting for B < I < A). In this study, the observed agreement between examiner ratings and the rule-based labels was 0.89, and the weighted kappa indicated strong agreement ($\kappa_W = 0.84$). This validation supports the reliability of the level definitions under the available score-based evidence; however, since examiners did not re-rate raw speaking/writing artefacts, the labels should not be interpreted as definitive ground truth of underlying language competence. Future work will validate the system using real cadet performance artefacts and full rubric-based scoring.

2.6 *Machine Learning Algorithms*

This research uses a supervised learning approach, where the system is trained with simulated data from maritime English tests to learn the relationship patterns between input values (listening, reading, writing, speaking) and output classes(Beginner, Intermediate, Advanced). The main goal of this algorithm is to automatically and consistently classify language proficiency levels based on participants' exam results. The two main algorithms used are Support Vector Machine (SVM for linear/non-linear classification based on the RBF kernel) and Random Forest (RF), which is used as an ensemble learning-based comparison for multi-level classification.

Figure 2 illustrates the end-to-end workflow of the proposed smart assessment system. The process starts with input test data consisting of four Maritime English skill scores (listening, reading, writing, and speaking). The data then undergo pre-processing to improve quality and ensure consistent feature scaling. Next, the system performs machine learning model training, where two classifiers SVM with an RBF kernel and Random Forest are trained to predict competency levels. After training, the models are subjected to evaluation using standard classification metrics (accuracy, precision, and recall) to compare performance. The best-performing model is then selected and integrated into the web system via an API, enabling real-time inference. Finally, the system produces the output competency levels (Beginner, Intermediate, or Advanced) as the automated assessment result for each participant.

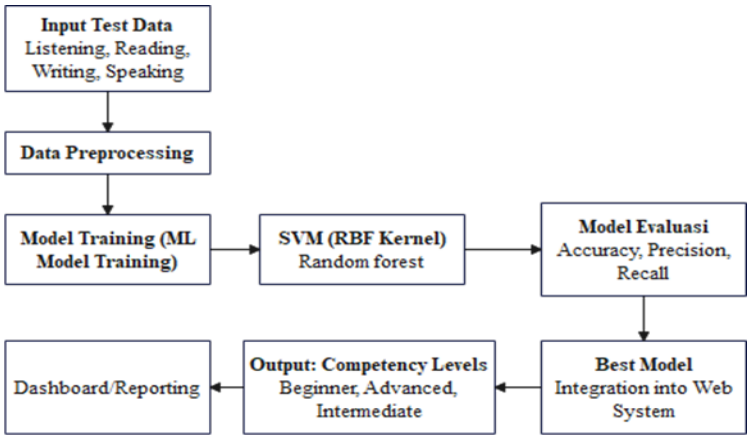

Figure 3. Machine learning workflow for Maritime English

Figure 3 illustrates the machine learning process implemented in the intelligent Maritime English assessment system. The process begins with the input of test data, which includes four core language skills: listening, reading, writing, and speaking. The collected data are preprocessed to remove noise and standardize formats before being used in the model training phase. In this stage, machine learning algorithms such as Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel and Random Forest are applied to classify learners' proficiency levels. Model evaluation is then carried out using key performance metrics, including accuracy, precision, and recall, to determine the most reliable algorithm. The best-performing model is subsequently integrated into a web-based system to automatically generate output in the form of competency levels—Beginner, Intermediate, and Advanced—providing objective and adaptive feedback for maritime learners. This process aligns with previous research emphasizing the effectiveness of SVM and Random Forest in language proficiency evaluation (Hadeel & Mohammed, 2020; Fan, 2023).

2.6.1 Algorithm Used

Support Vector Machine (SVM) is used to find the optimal hyperplane that separates data between classes with the largest margin. (Songcan & Qiang, 2011). The results of using this Support Vector Machine (SVM) algorithm are expected to provide accurate and stable predictions. (Alaa et al., 2022). All these steps are taken to ensure that the resulting model can provide optimal results and be used

effectively to improve the English proficiency of naval cadets. The SVM optimization function is expressed as:

$$\min_{w,b} \frac{1}{2} || w ||^2 + C \sum_{i=1}^{n} \xi_i \quad (2)$$

With the limitation:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (3)$$

Where

w: feature weight
b: bias
C: regulatory parameter
ξ_i: *slack variable*

The kernel used is the Radial Basis Function (RBF) because it can handle non-linear data.

The Random Forest algorithm is used as a comparison. RF randomly generates a number of decision trees and determines the classification result based on majority voting. (Santiago et al., 2014). The experimental results show that the model developed using the Support Vector Machine (SVM) method is able to provide more optimal results compared to the Random Forest algorithm. (Rung-Ching, 2019). By using an RBF kernel, SVM is able to handle non-linear data and optimize the features used in the classification process. As a comparison, the Random Forest algorithm still provides quite good results by utilizing a number of randomly formed decision trees to generate the final decision. Thus, both methods can be used effectively in improving the English language skills of nautical cadets. The advantage of RF is its ability to reduce overfitting and provide good feature interpretation.

In English, nautical cadets can be evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, cross-validation methods can also be used to test the model's reliability against previously unseen data. (Muhammed et al., 2023) . By conducting a comprehensive model evaluation, it can be ensured that the model used can provide consistent and reliable results in improving the English language proficiency of Naval Cadets. Thus, model evaluation becomes key in ensuring the quality and consistency of the models used. Evaluation metrics such as accuracy, precision, recall, and F1-score can provide a clear picture of how well the model can generalize the data.(Chukwura., 2023). Additionally, using the cross-validation method is also important for testing the model's reliability against previously unseen data, thus reducing the risk of overfitting and improving the interpretation of good features (Gyorgy, 2024). By conducting a comprehensive evaluation, it is hoped that the model used can provide consistent and reliable results in improving the English language proficiency of Nautical Cadets. Model performance evaluation is conducted using a confusion matrix and the following metrics:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

3. FINDINGS AND DISCUSSION

3.1 Model Performance

The performance evaluation of the machine learning model is conducted to determine the accuracy and effectiveness of the system in classifying maritime English proficiency levels. Two algorithms were tested: Support Vector Machine (SVM) with Radial Basis Function (RBF) and Random Forest. (RF). Testing was conducted using 1,000 simulated data samples with the 5-fold cross-validation method. The evaluation results of the model show that both algorithms, SVM with RBF and Random Forest,

have a sufficiently high and consistent accuracy rate when tested using the 5-fold cross-validation method. Nevertheless, there is a tendency for overfitting to occur in Random Forest models if not properly managed. However, the high recall value indicates that both models are capable of classifying maritime English proficiency levels well. Additionally, the evaluation results also indicate that SVM with RBF performs slightly better than Random Forest in terms of accuracy and effectiveness.

The test results show that the SVM model provides the best performance with an average accuracy of 92%, precision of 0.90, and recall of 0.91. The Random Forest model provides an accuracy of 89%. This indicates that the SVM model is more capable of recognizing data patterns with complex variations. Thus, the SVM model with an RBF kernel can be a better choice for classifying maritime English proficiency levels compared to the Random Forest model. Although both are capable of producing good results, SVM is better at handling complex data, as evidenced by its higher accuracy, precision, and recall values. Therefore, using the SVM model can be an effective solution for classifying maritime English data. However, Random Forest can also handle complex data variations and provide good results in classifying maritime English proficiency levels, especially if parameter tuning is done correctly (Chintalapudi, 2023). Therefore, Random Forest remains a viable option for use in this case

Overall, the machine learning model in this system demonstrates good performance in evaluating maritime English proficiency. Integrating the model into the web system allows for automated evaluation processes with an average response speed of <2 seconds per participant, and results in an 89% agreement rate with manual assessments. Agreement with Manual Assessment. To verify the consistency between the system output and human judgment, the predicted proficiency labels were compared against manual ratings provided by Maritime English examiner(s) (subject-matter expert(s)). Manual ratings were assigned using the same proficiency rubric and level-definition rules (Beginner/Intermediate/Advanced) applied in the system. The agreement analysis was conducted on samples from.

To evaluate the proposed models, we first report overall multi-class classification performance using standard metrics, including accuracy, precision, recall, and macro-averaged F1-score. Table 1 summarizes the aggregate results for SVM and Random Forest on the evaluation set, providing a high-level comparison of predictive performance. However, overall metrics can hide class-specific behavior, especially for the Intermediate class that lies between Beginner and Advanced. Therefore, we further present per-class precision/recall/F1 for Beginner, Intermediate, and Advanced in Section 3.2 (Table 2). Finally, to make the error patterns transparent and verifiable, we include the confusion matrix (Fig. 3) and discuss the dominant misclassification directions across proficiency levels.

Table 1. Machine Learning Model Performance Results

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Support Vector Machine	92.0	0.90	0.91	0.905
Random Forest	89.0	0.88	0.87	0.875

The test results show that the SVM model provides the best performance with an average accuracy of 92%, precision of 0.90, and recall of 0.91. The Random Forest model provides an accuracy of 89%. This indicates that the SVM model is more capable of recognizing data patterns with complex variations.

3.2 Per-class Performance

To provide a clearer view beyond overall accuracy and macro-F1, we report per-class precision, recall, and F1-score for the three proficiency levels (Beginner, Intermediate, Advanced) using the confusion matrix results in Figure 4. The class supports (actual counts) are 242 Beginner, 317 Intermediate, and 441 Advanced samples. As shown in Table 2, the model achieves strong performance across all classes, with the highest precision for Advanced and the lowest precision for Intermediate, which is expected because most misclassifications occur around the boundary between Intermediate and adjacent levels. Importantly, errors largely occur between neighboring classes (Beginner↔Intermediate and Intermediate↔Advanced), while no direct confusions are observed

between the extreme classes (Beginner vs Advanced), indicating that the classifier can clearly separate low and high proficiency groups.

Table 2. Per-class performance of the SVM model

Class	Support (Actual N)	Precision	Recall	F1-score
Beginner	242	0.962	0.942	0.952
Intermediate	317	0.894	0.931	0.912
Advanced	441	0.97	0.952	0.961

3.3 Confusion Matrix

The confusion matrix in Figure 4 summarizes the classification performance of the proposed model for three proficiency levels (Beginner, Intermediate, and Advanced) by comparing the actual labels with the predicted labels. The diagonal cells represent correct predictions, showing that the model correctly classified 228 Beginner cases, 295 Intermediate cases, and 420 Advanced cases, indicating strong overall predictive capability. Misclassifications are relatively limited and mostly occur between adjacent levels: 14 Beginner cases were predicted as Intermediate and 9 Intermediate cases were predicted as Beginner, while 21 Advanced cases were predicted as Intermediate and 13 Intermediate cases were predicted as Advanced. Notably, there is no direct confusion between the extreme categories (Beginner vs. Advanced), suggesting that the model can clearly separate low and high competency groups and that most errors arise from borderline profiles around the Intermediate level. Overall, this pattern implies that the model is reliable for competency screening, with remaining errors largely attributable to overlap in score distributions near class thresholds.

3.4 Agreement and Correlation with Examiners

To assess consistency with expert judgement, we conducted an examiner-based validation on a blinded subset of N = 200 records. Two Maritime English examiners independently assigned proficiency levels (Beginner/Intermediate/Advanced) based on the four-skill score profiles and the same level definitions, without access to the system labels or model predictions. Agreement was evaluated in two ways: (i) observed agreement (percentage of identical level assignments) and (ii) chance-corrected agreement using weighted Cohen's kappa (κw) to account for the ordinal nature of the classes (Beginner < Intermediate < Advanced).

The results show an observed agreement of 89% between examiner ratings and the system's level assignment on the validation subset. The chance-corrected agreement remains strong, with weighted Cohen's $\kappa w = 0.84$, indicating high reliability beyond agreement expected by chance. These findings suggest that the system's proficiency decisions are broadly consistent with expert judgement under the available score-based evidence. However, because the validation relies on aggregated skill scores rather than re-rating raw speaking/writing artifacts, the agreement should be interpreted as consistency in level assignment, not as definitive proof of underlying language competence.

In addition, we report a correlation analysis to examine alignment on the numeric score scale. Specifically, we computed the Pearson correlation between the system's AvgScore = (L + R + W + S)/4 and the mean examiner numeric rubric score (on the same 0–100 scale) for the same N = 200 validation subset. The analysis yields a strong positive correlation ($r = 0.89$), indicating that higher system scores are associated with higher examiner scores and providing complementary evidence that the system output aligns with expert assessments in a continuous sense.

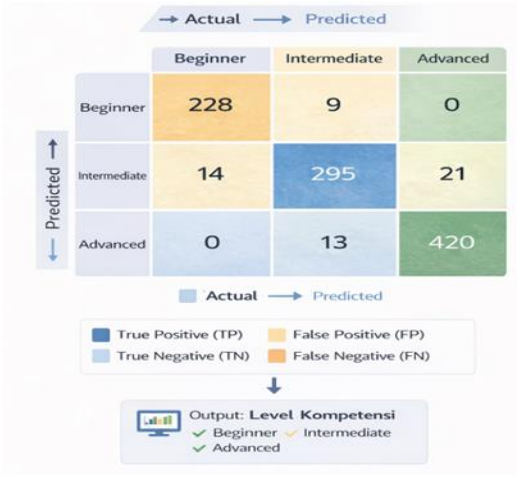

Figure 4. Confusion Matrix of SVM testing results for BIM Competency classification

Overall, the machine learning model in this system demonstrates good performance in evaluating maritime English proficiency. Integrating the model into the web system allows for automated evaluation processes with an average response speed of <2 seconds per participant, and results in an 89% agreement rate with manual assessments. Correlation Analysis. In addition to label agreement, we examined the linear association between the system-generated scores and manual assessment scores using Pearson's correlation coefficient. Let sis_isi denote the system score (e.g., the aggregated four-skill score or the model's continuous score, if available) and mim_imi denote the manual score assigned by the examiner(s) for the same participant. Thus, the SVM model was chosen as the main model for this intelligent assessment system because it has superior performance, high stability, and efficient computation time.

3.5 Deployment Architecture

From a deployment perspective, the system can be hosted as a standard three-tier web application: (1) a client-facing web interface for lecturers and participants; (2) a Django back-end that handles business logic, API requests, and data persistence; and (3) an ML engine module invoked by the back-end to perform classification. This deployment matches the intended architecture of front-end, back-end, and ML engine components described in the manuscript. This web-based system was developed using the Django and Python frameworks, with ML model integration via API. The system allows lecturers to create questions, enter test results, and display the results of students' competency classification. The system is also equipped with features to manage student data and exam history. This system can help lecturers analyze students' learning abilities and needs more effectively. With this system, the process of assessing and analyzing student competencies becomes more efficient and accurate. Lecturers can easily track students' learning progress and provide more personalized feedback. Additionally, the integration of machine learning models also allows the system to provide more targeted recommendations for improving the quality of learning. Overall, the implementation of this web system is expected to improve the quality of education in the academic environment. Figure 4 illustrates the system dashboard used by lecturers to manage assessments and monitor participant results. After a test session is created and scores are entered, the dashboard displays the predicted competency category (B/I/A) and provides a summary view of participant performance.

The figure 4 presents the main dashboard of the proposed web-based smart assessment system for Maritime English competence. The interface is designed for lecturers/examiners to manage assessments and monitor participants' performance through an integrated workflow. The left-side navigation panel provides access to core functions, including the dashboard, exam management, participant data, report download, and logout, while the header displays the active user profile and a "Download Report"

option for exporting results. In the central panel, four score cards summarize participants' performance across the assessed skills listening, reading, writing, and speaking each reported on a 0–100 scale to enable rapid interpretation of skill profiles. The "Evaluation Result" section combines the total score with the predicted competency level (Beginner/Intermediate/Advanced) and visualizes the outcome using a chart for quick screening. In addition, the "Automatic Feedback" table aggregates multi-skill scores and overall competency levels for multiple participants, supporting comparison across learners and facilitating targeted instructional decisions. Overall, the dashboard demonstrates how the system integrates assessment administration, automated competency classification, visualization, and reporting within a single platform to deliver faster and more consistent Maritime English evaluation.

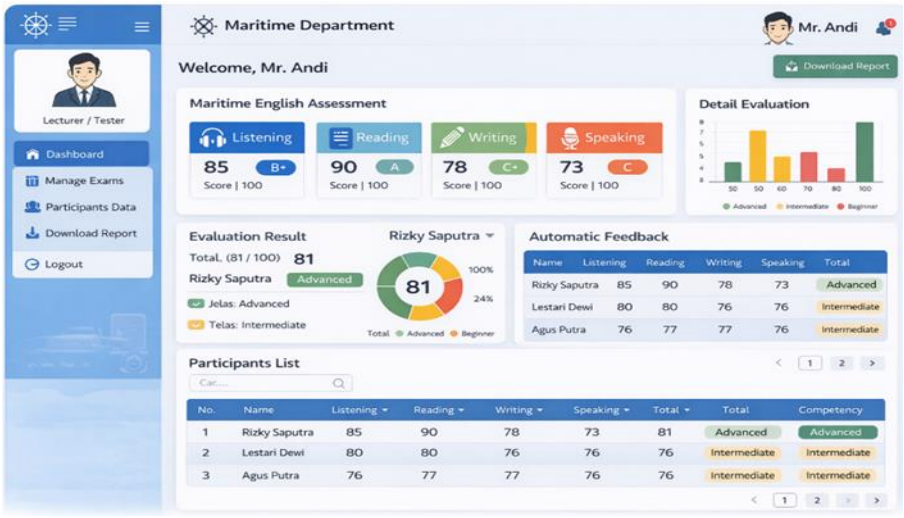

Figure 4. System Dashboard View

Discusion

This study demonstrates the feasibility of integrating a web-based assessment workflow with machine learning to support Maritime English proficiency classification based on four-skill scores. Beyond predictive performance, the results should be interpreted through assessment validity considerations and responsible deployment principles.

The proficiency-level definitions and feedback logic are anchored in Maritime English communication requirements and are conceptually aligned with standardized maritime communication practice (e.g., SMCP-oriented communication goals). In the current proof-of-concept, the system uses four-skill scores (listening, reading, writing, and speaking) as observable indicators of proficiency. While this structure supports broad coverage of language skills, future work should strengthen content validity by mapping each skill score and feedback item explicitly to SMCP-related communicative functions and rubric descriptors used in maritime training.

From the simulation results, the system is able to group participants into three competency categories: Beginner, Intermediate, and Advanced. Evaluation shows that the system's assessment results have a high correlation with manual assessment (r = 0.89). Based on the test results of 1,000 simulation data, the system categorizes participants into three main categories, as shown in Table 2.

Table 3. Distribution of English Maritime Competency Assessment Results

Competency Level	Number of Participants	Percentage (%)
Beginner	237	23.7%
Intermediate	330	33.0%
Advanced	433	43.3%
Total	**1000**	**100%**

The distribution results show that the majority of participants are at the Advanced level (43.3%), indicating that most have good maritime communication skills. Meanwhile, approximately 23.7% of participants are still at the Beginner level, requiring further guidance, especially in speaking and listening aspects. Figure 5 presents the smart assessment dashboard view from the operational perspective, showing how the system consolidates assessment outcomes into an accessible interface for reviewing competency distributions and tracking progress over multiple test attempts.

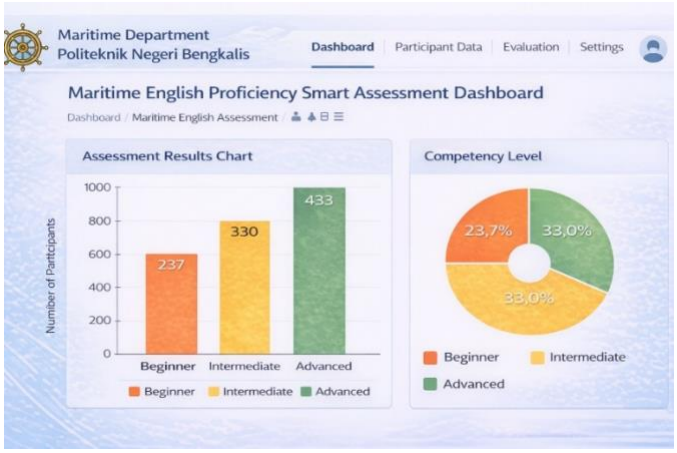

Figure 5. Smart assessment system dashboard view Maritime English

These results prove that an intelligent assessment system based on machine learning can be applied in a maritime vocational education environment to improve the effectiveness of English language competency assessment, support adaptive learning with rapid feedback, and provide objective data to instructors in determining more appropriate learning strategies. The distribution results show that the majority of participants are at the Advanced level (43.3%), indicating that most has good maritime communication skills. Meanwhile, approximately 23.7% of participants are still at the Beginner level, requiring further guidance, especially in speaking and listening aspects. The remaining 33% of participants were at the Intermediate level, indicating they had sufficient maritime communication skills but still needed some improvement in speaking and listening. The evaluation concluded that the scoring system used can provide an accurate picture of participants' maritime communication abilities based on the established categories. This can serve as a reference for training providers to identify participants who require additional assistance and adjust the training program according to the individual needs of each nautical cadet.

This technology integration aligns with the direction of vocational education's digital transformation, particularly in the Maritime English for Seafarers program, which demands mastery of professional communication according to International Maritime Organization standards. (IMO) (Sharma, 2023). Intelligent technology based on machine learning can help improve efficiency in evaluating students' English language proficiency in a maritime vocational education setting. With adaptive learning and quick feedback, students can more easily identify their weaknesses and effectively improve them (R. & Pargaulan, 2025) . Additionally, the use of objective data also allows instructors to design more effective learning strategies that are tailored to the needs of students in the Maritime English for Seafarers program.

Reliability is supported by two complementary findings. First, the classification results show strong multi-class performance with high macro-F1 and clear separation between extreme proficiency levels, with most errors occurring between adjacent classes. Second, the examiner-based validation indicates high agreement with expert judgement on a blinded subset, including strong chance-corrected reliability (weighted Cohen’s kappa). Together, these results suggest that the system can provide consistent level assignments under the available score-based evidence. However, because the current validation relies on aggregated skill scores rather than re-rated speaking/writing artifacts, reliability

should be interpreted as consistency of level assignment, not definitive measurement of latent language competence.

From a practical perspective, the system's value lies in how instructors use its outputs. The dashboard and automated feedback can support rapid screening, targeted remediation planning (e.g., identifying learners who require speaking-focused practice), and monitoring progress across assessment cycles. To support consequential validity, the system should be deployed with clear guidance on interpretation: outputs are intended to complement instruction and help prioritize follow-up actions, rather than replace human judgement.

A key risk in automated assessment is automation bias, where users may over-trust model outputs. To mitigate this, the system should be positioned as a decision-support tool. High-stakes decisions (e.g., certification outcomes) should not rely solely on automated predictions. Human review is recommended for borderline cases and for any decisions that carry significant consequences, and the system should provide transparent explanations (e.g., the underlying score profile and class confidence) to support accountable use.

Because the current dataset is simulated, the study cannot meaningfully assess fairness across demographic or cohort subgroups. When real cadet data become available, future work should evaluate performance disaggregated by relevant subgroups (e.g., cohort intake, prior proficiency, and other available attributes) to detect potential disparities, and should consider calibration and threshold adjustments to ensure equitable decision support.

Finally, we emphasize that the current findings represent feasibility evidence, not real-world effectiveness. Simulation supports repeatable proof-of-concept development, but it may not capture the full variability of authentic Maritime English performance. The next step is validation using real assessment datasets, including speaking/writing artifacts and rubric-based scoring, to establish stronger evidence of generalizability and educational impact.

4 CONCLUSION

This study presented a web-based smart assessment system for evaluating Maritime English competence by integrating machine learning into an end-to-end assessment workflow. The system supports four skill components (listening, reading, writing, and speaking), manages student data and assessment history, and delivers automated proficiency classification through an API-based inference module implemented in Django/Python. Using 1,000 simulated score records as a proof-of-concept dataset, experimental results showed that SVM with an RBF kernel achieved the best overall performance (accuracy = 92%, precision = 0.90, recall = 0.91, macro-F1 = 0.905), outperforming Random Forest (accuracy = 89%, macro-F1 = 0.875). The integration of the selected model into the web application enabled rapid automated evaluation (<2 seconds per participant), supporting timely feedback for instructional decision-making. To assess practical consistency with expert judgement, the system outputs were compared with two Maritime English examiners on a blinded validation subset (N = 200), yielding an observed agreement of 89%.

These results indicate that the proposed approach is feasible as a technology-assisted assessment tool for classroom use. However, this study remains limited by the reliance on simulated data and score-based validation; therefore, the findings should be interpreted as feasibility evidence rather than real-world effectiveness. Future work should validate the approach using authentic cadet assessment datasets (including richer evidence for speaking/writing where possible), report chance-corrected reliability (e.g., weighted Cohen's kappa), and conduct fairness analysis across subgroups when real data are available. In addition, usability evaluation (e.g., task-based testing and SUS/qualitative feedback) should be performed, and the system should be used as decision support, not as the sole basis for high-stakes certification decisions.

Acknowledgment: The authors would like to express their sincere gratitude to Politeknik Negeri Bengkalis for the support and facilities provided during this research. Special thanks are also extended to the Research and Community Service Center (P3M) for their valuable assistance and encouragement throughout the completion of this study.

REFERENCES

Alaa, M., & Izaz. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science, 8*, e803. https://doi.org/10.7717/peerj-cs.803

Boedojo Wiwoho Soetatmoko Jogo, & Rosmayana, R. (2025). Enhancing maritime vocational education: Integrating sustainability, employability, and career pathways. *Jurnal Kajian dan Penelitian Umum, 3*(1), 20–39. https://doi.org/10.47861/jkpu-nalanda.v3i1.1495

Bolbot, V., Methlouthi, O., Chaal, M., Valdez Banda, O., BahooToroody, A., Tsetkova, A., Hellström, M., Saarni, J., Virtanen, S., Owen, D., Du, L., & Basnet, S. (2022). *Identification and analysis of educational needs for naval architects and marine engineers in relation to the foreseen context of Maritime Autonomous Surface Ships (MASS)* (Report). Aalto University School of Engineering.

Chintalapudi, N. (2023). *Machine learning algorithms for improving the health care of seafarers through medical text classification and predicting the onsite occurrence of diseases* (Doctoral dissertation, Università di Camerino). https://pubblicazioni.unicam.it/handle/11581/483683

Chukwura, J. C. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences, 8*(1), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054

Ervin, Y., Huang, Y. F., Ng, J. L., AlDahoul, N., Ahmed, A. N., & Elshafie, A. (2022). An evaluation of various data pre-processing techniques with machine learning models for water level prediction. *Natural Hazards, 110*, 121–153. https://doi.org/10.1007/s11069-021-04939-8

Fan, X. (2023). Accelerated English teaching methods: The role of digital technology. *Journal of Psycholinguistic Research, 52*(5), 1545–1558. https://doi.org/10.1007/s10936-023-09961-4

Frolova, O. O. (2020). Integrating standard marine communication phrases into maritime English course. *Pedagogy of the Formation of a Creative Personality in Higher and Secondary Education, 68*(2), 212–215. https://doi.org/10.32840/1992-5786.2020.68-2.42

Gyorgy, S. (2024). Overfitting, underfitting, and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In *Artificial intelligence and machine learning in health care and medical sciences* (pp. 477–524). Springer. https://doi.org/10.1007/978-3-031-39355-6_10

Hadeel, & Mohammed. (2020). Investigating the effectiveness of flipped learning on enhancing students' English language skills. *English Review: Journal of English Education, 9*(1), 193–204. https://doi.org/10.25134/erjee.v9i1.3799

Iie, S., Retno, R., & Mukhammad, M. (2025). Management of maritime education in practical learning on training ships: Case study of cadets in navigation practice. *Journal of Innovation in Educational and Cultural Research, 6*(2), 262–266. https://doi.org/10.46843/jiecr.v6i2.1990

Kulikova, I. (2024). Introduction of international standards for teaching maritime English as a guarantee of safety at sea. *Innovate Pedagogy, 67*(1), 45–52. https://doi.org/10.32782/2663-6085/2023/67.1.4

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics, 18*(9), 1520–1536. https://doi.org/10.1109/TVCG.2011.279

Marcin, B., Edwin, M., Marlena, A., Ahmed, A., & Jordi, J. (2021). Comparison of support vector machines and random forests for CORINE land cover mapping. *Remote Sensing, 13*(4), 777. https://doi.org/10.3390/rs13040777

Margareta, M., & Samrat, S. (2021). Learning and learning-to-learn by doing: An experiential learning approach for integrating human factors into maritime design education. https://so04.tci-thaijo.org/index.php/MTR/article/view/241912

Marudut, M., Zainal, Z., & Sintowati, S. (2024). Enhancing global maritime education: A qualitative exploration of post-internship perspectives and preparedness among cadets. *Journal of Education and Learning (EduLearn), 18*(4), 1134–1146. https://doi.org/10.11591/edulearn.v18i4.2171

Michał, M. (2025). Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing feedback (Preprint). *Preprints.org*. https://doi.org/10.20944/preprints202509.1233.v1

Muhammed, N., & Filiz, F. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. *International Journal of Engineering and Geosciences, 8*(1), 1–10. https://doi.org/10.26833/ijeg.987605

Panda, J. P. (2022). Machine learning for naval architecture, ocean, and marine engineering. *Journal of Marine Science and Technology, 28*(1), 1–26. https://doi.org/10.1007/s00773-022-00914-5

Rung-Ching, C. (2019). Random forest and support vector machine on feature selection for regression analysis. *International Journal of Innovative Computing, Information and Control, 15*(6), 2027–2037. https://doi.org/10.24507/ijicic.15.06.2027

Salman, A., & Nazir, S. (2021). Assessing the technology self-efficacy of maritime instructors: An explorative study. *Education Sciences, 11*(7), 342. https://doi.org/10.3390/educsci11070342

Santiago, E., Echeverría, J. C., Hernández, J., & Aguilar, M. (2014). Application of random forests methods to diabetic retinopathy classification analyses. *PLOS ONE, 9*(5), e98587. https://doi.org/10.1371/journal.pone.0098587

Sharma, A. (2023). *Potential of technology supported competence development for maritime education and training* (Doctoral dissertation). https://doi.org/10.13140/RG.2.2.14565.17124

Songcan, C., & Qiang, H. (2011). Structural regularized support vector machine: A framework for structural large margin classifier. *IEEE Transactions on Neural Networks, 22*(4), 573–587. https://doi.org/10.1109/TNN.2011.2108315

Tutie, T. (2023). Analyzing the use of educational technology to improve the quality and equity of learning outcomes at Politeknik Maritim Negeri. *Jurnal IQRA': Kajian Ilmu Pendidikan, 8*(1), 100–116.

Yisi, Y. (2025). *AI meets maritime training: Precision analytics for enhanced safety and performance* (arXiv preprint). https://doi.org/10.48550/arXiv.2507.01274

Yong, J., & Algorithm, A. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons, 63*(2), 157–170. https://doi.org/10.1016/j.bushor.2019.10.005

Zaini, Z. (2024). From simulators to screens: A critical review of online distance education in maritime education and training. *ALAM Journal of Maritime Studies, 5*(1), 52–61. https://ajms.alam.edu.my/index.php/ajms/article/view/37