Vol. 8, No. 2, April 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

A Multi-Algorithm Approach for Predicting OSCE Exam Passing Status

Zulkifli[1], Panji Bintoro[2], Fitriana [3], Muhammad Galih Ramaputra[4], Hafsah Mukaromah[5]

[1]Department of Informatics Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia

[2]Department of Software Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia

[3]Department of Midwifery, Faculty of Health, Aisyah University, Indonesia

[4]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung, Lampung, Indonesia

[5] Department of Pharmacy, Faculty of Health, Aisyah University, Indonesia

Received:
October 12, 2025
Revised:
March 3, 2026
Accepted:
March 27, 2026
Published:
April 12, 2026

Corresponding Author:
Author Name*:
Zulkifli Zulkifli
Email*:
zulkifli@aisyahuniversity.ac.id

DOI:
10.63158/journalisi.v8i2.1518

Abstract. This study proposes a digital decision support system to automate Objective Structured Clinical Examination (OSCE) evaluation using machine learning. Five algorithms were experimentally compared: Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (kNN). The model was developed using a dataset of 439 clinical competency records from midwifery students at Aisyah Pringsewu University. Eight clinical skill variables were used as input features, including baby massage, newborn care, and family planning services. Model performance was validated using 5-fold cross-validation to ensure stability and reliability. The results showed that all algorithms achieved strong performance, with accuracy values above 90%. Among them, SVM produced the best result, reaching 100% classification accuracy, while Random Forest and Neural Network also demonstrated high effectiveness, with the Neural Network achieving an average validation accuracy of 95%. Although the findings are promising, the model is limited by the relatively small dataset and its focus on a specific educational context. Therefore, further refinement and broader validation are needed before implementation in other institutions. This study provides a foundation for developing digital assessment systems that can help educators identify students who need additional support before final examinations.

Keywords: OSCE Assessment, Machine Learning, Classification, Decision Support System, Student Performance Prediction

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

A standardized competency test called the Objective Structured Clinical Examination (OSCE) is used to evaluate health students' eligibility for graduation [1]. It is administered before to the professional oath and uses a fixed passing score to judge whether or not students are qualified. Those who don't pass must go through remediation. OSCE is widely used in many health-related fields, such as medicine, nursing, midwifery, and pharmacy [2]. It usually consists of brief simulated clinical scenarios that evaluate abilities like patient management, communication, and clinical reasoning [3]. The OSCE has been included into national medical certification programs in a few nations, including France. The OSCE for midwifery students is administered at the institutional level through a number of timed stations that assess particular practical competencies [4]. On our campus, the OSCE exam for midwifery students includes 8 consecutive 7-minute stations, consisting of tests in areas of expertise including family planning skills, realization, prenium, octitocin, pregnancy gymnastics, newborn handling, endorphin, and infant massage.

The OSCE graduation rates remain relatively low, with 64.38% for DIII Nursing, 71.78% for DIII Midwifery, and 53.61% for the Ners profession, indicating that many students fail to pass the competency test and may experience decreased motivation [5] [6]. To address this issue and improve pass rates across health disciplines, there is a need to develop tools and predictive models that can estimate students' likelihood of passing the OSCE, providing an early reference or simulation before they take the actual exam [7].

Some research that discusses osce is research conducted by Tufayl in 2021, in his research the researcher designed and implemented an online osce assessment using zoom tools and this research was conducted during COVID-19 [8]. In developing tools and models for predicting the passing status of the OSCE test, the author conducts a literature study of research related to the development of multi-algorithm applications, data mining and machine learning. One of them was conducted by Feda Anisah Makkiyah in 2024, in this study explaining the development of applications and the application of multi-algorithms for predicting diabetes status [7].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

This research will be developed experimentally, by producing tools or prediction models to measure the passing status of the osce test using several algorithms, including NB, SVM, RF, kNN, and NN. These tools or models can be used to predict osce test pass status. There are several related studies on osce test and the use of multi-algorithm as a model for dataset prediction. One of them was conducted by Tufayl in 2021, in his research the researcher designed and implemented an online osce assessment using the zoom tool and this research was conducted during COVID-19 [8]. In developing tools or prediction models for passing the osce test, the author conducts a literature study of research related to the development of multi-algorithm applications, data mining and machine learning. One of them was conducted by Feda Anisah Makkiyah in 2024, in this study explaining the development of applications and the application of multi-algorithms for predicting diabetes status [7]. Furthermore, research using machine learning that January F. Naga in 2024, in his research conducted Deciphering Digital Discourse: Detecting Cyber bullying Patterns in Filipino Tweets Using Machine Learning [9].

From existing studies, it is only limited to the use of tools or platforms such as zoom or osce test models, not yet developed to a practical stage, where tools or models developed in the form of this application can be used by health students in preparation before the osce test, because with these tools or models it can be used as an initial reference or try out by involving 8 types of skills assessment, namely family planning skills, awareness, prenium, octitocin, pregnant gymnastics, handling newborns, endorphin, and baby massage. Then this tool can also be used by lecturers as a digital assessment tool so that graduation results can be known quickly and precisely.

The majority of previous research has focused on the implementation of the Objective Structured Clinical Examination (OSCE) [10], student perceptions, learning effectiveness, or the development of digital and AI-assisted assessment tools [11], despite the OSCE's widespread use as a standardized method for evaluating clinical competencies. Despite recent research investigating the use of AI to enhance OSCE evaluation or training environments, predictive analytics has not yet been widely used to evaluate students' chances of passing the OSCE [12]. Furthermore, most current research focuses on correlations between exam components or uses traditional statistical techniques to analyze OSCE outcomes rather than creating predictive models that could help students and instructors make early decisions [13].

Predictive frameworks that apply machine learning for early OSCE performance assessment are critically lacking, despite the literature currently in publication concentrating on student impressions and conventional statistics. In order to close this gap, five machine learning algorithms (SVM, NN, RF, kNN, and Naive Bayes) are experimentally evaluated in this work to determine the best method for projecting graduation outcomes and offering educational institutions a decision-support tool. The study compares the predictive performance of these algorithms in an effort to identify the optimal approach for estimating OSCE passing outcomes. Additionally, it aims to offer a decision-support tool that can assist educational institutions in identifying students who may require additional study time prior to an exam.

2. METHODS

The steps for creating tools and prediction models to pass the OSCE test are NB, kNN RF, SVM, and NN algorithms are used to measure the accuracy level.

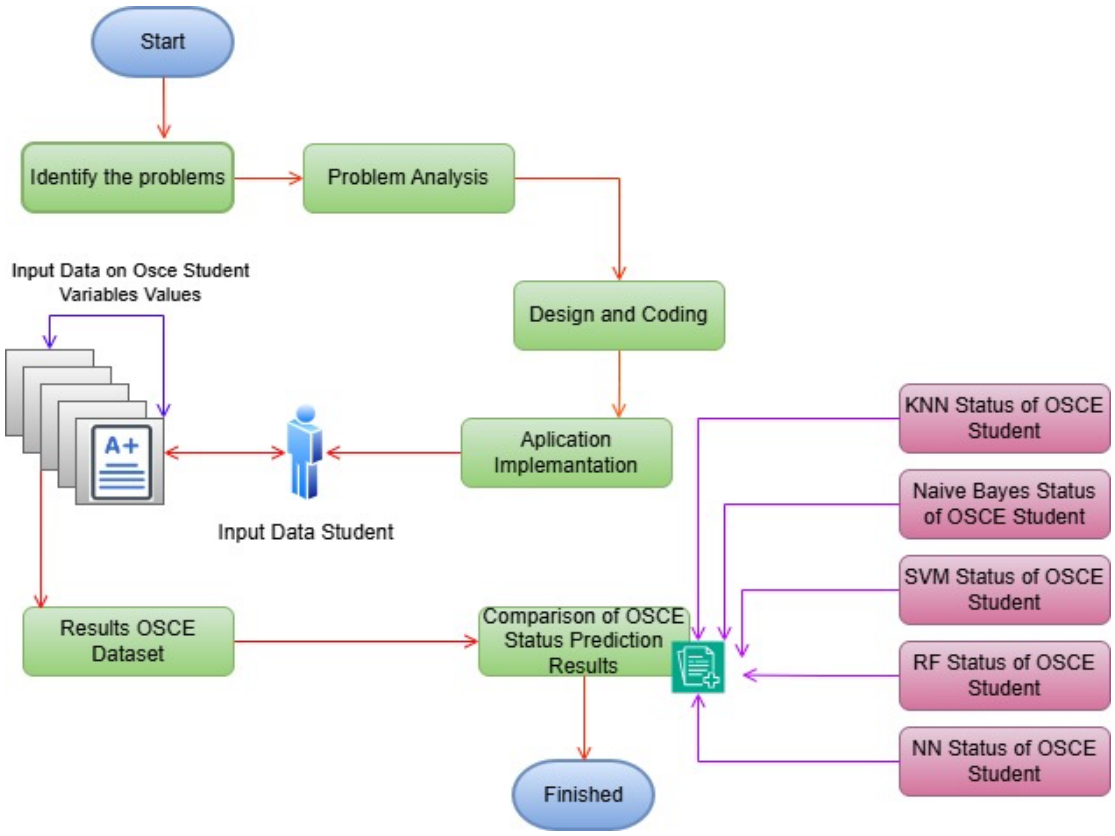

Figure 1. The Stages of Developing Tools and Models for Predicting the Passing Status of the OSCE Test by Measuring the Accuracy of Passing

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The process of creating a prediction model employing many machine learning algorithms to ascertain students' OSCE status is depicted in Figure 1. In order to identify obstacles in predicting students' performance on the Objective Structured Clinical Examination (OSCE), the study starts with the problem identification step. This phase emphasizes the necessity of a predictive system that can help teachers identify kids who might need more academic support prior to the test. Problem analysis is the next step, where variables affecting students' OSCE performance are looked at. A structured dataset is created by identifying and organizing pertinent factors, including academic scores, practical exams, attendance, and prior academic achievement. To make sure the dataset is appropriate for machine learning research, the obtained data are then processed through data preparation, which includes cleaning and normalization. After that, the design and coding phase concentrates on creating the predictive system by putting machine learning algorithms into practice and setting up the computational workflow. The technology is subsequently put to use during the application installation phase, where predictions of OSCE results are produced by processing student data. Several methods, such as SVM, NN, RF, kNN, and NB, are used to assess predicting performance. An OSCE outcomes dataset is created by compiling the prediction results produced by each algorithm. To choose the best model, these outcomes are then evaluated using assessment parameters like accuracy and other categorization metrics. The best-performing algorithm for predicting OSCE status is finally identified in the study's conclusion, offering educational institutions a helpful decision-support tool to enhance early identification of students' OSCE exam readiness.

An application-based system structure for handling and evaluating student data from the Objective Structured Clinical Examination (OSCE) is shown in Figure 2. Academically, the system flow starts at the data entry stage through a dashboard application, where different clinical metrics and student competencies like endorphins, self-awareness, and KB scores are gathered and combined into a structured dataset. The dataset includes 439 student observation reports gathered from Aisyah University's midwifery competency tests. There are 221 people who have passed the exam and 218 people who have failed it. After that, the data is exported in CSV format for additional processing, such as feature selection and model validation, in a cloud computing environment (Google Colab).

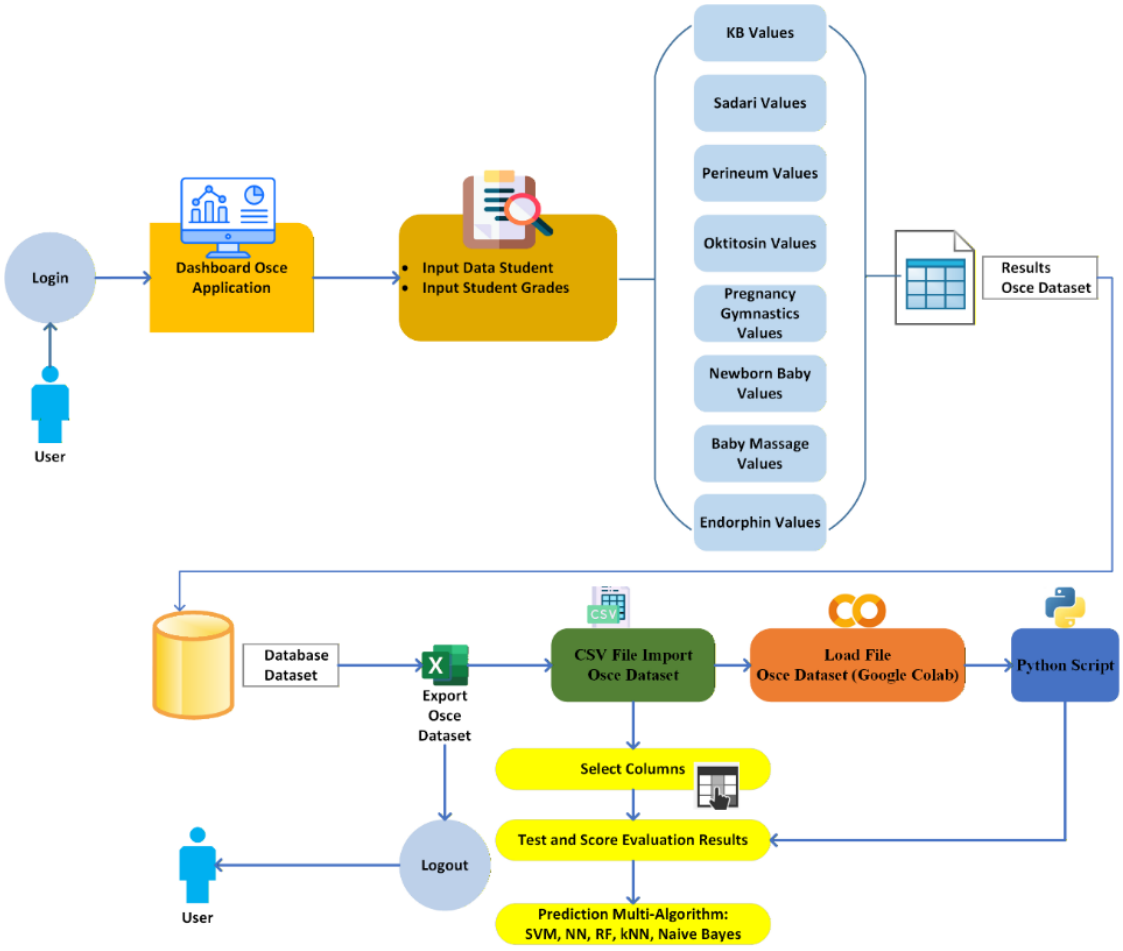

Figure 2. Research Framework for Measuring the Accuracy of OSCE Test Pass Status

The creation of a multi-algorithm prediction system, which concurrently uses NN, SVM, RF, kNN, and NB models to ascertain students' graduation status, is the apex of this framework. The 439 records were divided into five subsets using 5-Fold Cross-Validation to guarantee that each data point was used for both training and validation. To get the TP, TN, FP, and FN values, a Confusion Matrix was created for each fold (or an aggregate of all folds). The probability outputs of these cross-validation iterations were used to calculate the ROC Analysis and AUC values, which quantify the trade-off between sensitivity and specificity. This design enables complete automation from the data collection phase to the effective and integrated interpretation of prediction outcomes. In order to identify the most accurate predictor of OSCE results, the main contribution is a comparative performance analysis of five algorithms.

3. RESULTS AND DISCUSSION

The steps involved in creating models and tools to measure the accuracy of OSCE test pass status using NB, kNN RF, SVM, and NN algorithms. The need for a tool and model to measure the accuracy of passing the OSCE test, which involves eight skill assessment variables—the value of birth control, awareness, prenium, octitocin, pregnant gymnastics, handling newborns, endorphin, and baby massage—has been identified at this point.

3.1. Model Training

In this stage, problem identification will be carried out, including:

1) The need for tools and models to measure the accuracy of osce test pass status by measuring the accuracy of pass status, to be used in decision-making or policy implementation.
2) The developed model should be applicable to health services.
3) Assessing how well NB, kNN RF, SVM, and NN algorithms predict the success of an OSCE test.

3.2. Problem Analysis

Communicating with tools users to understand the software expected by users both students and lecturers by conducting interviews, and discussions.

3.3. Dataset Database

This dataset, which is displayed in Table 1, will be utilized to measure the prediction accuracy of the OSCE exam passing status using NB, kNN RF, SVM, and NN algorithms. A portion of the multivariate Objective Structured Clinical Examination (OSCE) dataset, which includes students' clinical competencies and academic data, is shown in Table 1. In terms of structure, this table combines 11 clinical performance variables specific values like KB, Sadari, Perineum, Oxytocin, pregnant exercise, infant care, baby massage, and endorphins with student identity elements (data code and academic year) and validation scores. The OSCE Permission property displays a status of "Y," whereas the target variable the graduation status (Passed Status)—dominantly displays the outcome "1" (passed). These clinical characteristics are used as input variables in a machine learning-based predictive model that uses this dataset as an empirical foundation to match competency patterns against student graduation outcomes. The dataset includes 439 student

observation reports gathered from Aisyah University's midwifery competency tests. There are 221 people who have passed the exam and 218 people who have failed it.

Table 1. OSCE Dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MHS-01	2020	20040359	Active	0	86	93	72	81	90	83	90	74	0	Y
MHS-02	2021	21040032	Active	74	71	70	75	70	76	70	79	73	0	Y
MHS-03	2021	21040140	Active	87	86	89	87	88	96	90	82	88	1	Y
MHS-04	2021	21040131	Active	93	82	93	87	91	93	88	87	89	1	Y
MHS-05	2021	21040454	Active	79	78	85	87	80	87	81	83	83	1	Y
MHS-06	2021	21040226	Active	96	97	98	96	97	98	98	96	97	1	Y
MHS-07	2021	21040227	Active	77	86	82	77	82	90	83	83	83	1	Y
MHS-08	2021	21040404	Active	89	86	92	90	81	92	88	90	89	1	Y
MHS-09	2021	21040120	Active	91	84	93	83	91	92	86	88	89	1	Y
MHS-010	2021	21040486	Active	96	89	94	96	95	10	90	91	83	1	Y

Note: 1=Code Data, 2=Year, 3= Exam Number, 4= Active Status, 5= KB Values, 6=Sadari Values, 7=Perineum Values, 8=Oktitosin Values, 9= Pregnancy Gymnastics Values, 10=Newborn Baby Values, 11=Baby Massage Values, 12= Endorphin Values, 13= Validation Values, 14= Passed Status, 15= Osce Permission

3.4. Dataset Normalization

The Min-Max Scaling (MinMaxScaler) approach, which converts all feature values into a uniform range between 0 and 1, is used to normalize the dataset shown in Table 2. By removing the impact of varying data sizes, this normalization procedure is crucial to ensuring that each variable contributes proportionately in ensuing analyses, especially in predictive modeling. Family planning (KB), SADARI, perineum care, oxytocin, pregnancy gymnastics, newborn care, infant massage, endorphin, and validation scores are among the competency variables represented by each of the ten samples (MHS-01 to MHS-010) in the dataset. After normalization, values near 0 indicate inferior performance in comparison to other samples within the same feature, whereas values closer to 1 suggest higher performance levels. While some samples, like MHS-01 and MHS-010, show lower values in particular skills, indicating areas that need improvement, MHS-06 consistently shows high values across all categories, indicating great overall proficiency. All things considered, using Min-Max normalization improves data comparability and facilitates the creation of more precise and objective models. A normalized sample dataset is displayed in Table 2.

Table 2. Dataset Normalization

1	2	3	4	5	6	7	8	9	10
MHS-01	0.00	0.88	0.94	0.73	0.82	0.91	0.84	0.91	0.76
MHS-02	0.75	0.73	0.71	0.76	0.71	0.77	0.71	0.80	0.75
MHS-03	0.88	0.88	0.90	0.88	0.89	0.97	0.91	0.83	0.90
MHS-04	0.94	0.84	0.94	0.88	0.92	0.94	0.89	0.88	0.91
MHS-05	0.80	0.80	0.86	0.88	0.81	0.88	0.82	0.84	0.85
MHS-06	0.97	1.00	1.00	0.97	0.98	1.00	1.00	0.97	1.00
MHS-07	0.78	0.88	0.83	0.78	0.83	0.91	0.84	0.84	0.85
MHS-08	0.90	0.88	0.93	0.91	0.82	0.93	0.89	0.91	0.91
MHS-09	0.92	0.86	0.94	0.84	0.92	0.93	0.87	0.89	0.91
MHS-010	0.97	0.91	0.95	0.97	0.96	0.10	0.91	0.92	0.85

Note: 1= Code Data, 2= KB Values, 3=Sadari Values, 4=Perineum Values, 5=Oktitosin Values, 6= Pregnancy Gymnastics Values, 7=Newborn Baby Values, 8=Baby Massage Values, 9= Endhorphin Values, 10= Validation Values

3.5. Training Architecture and Training Performance

All suggested models underwent cross-validation during the training phase with a K-Fold value of 5. The whole dataset was used to train each layer of the NN model using the Stochastic Gradient Descent (SGD) optimization algorithm, with parameter settings such as epoch=10, , and validation_split=0.2, and batch_size 32 [14], [15]. The NN design utilized in this investigation is depicted in Figure 3.

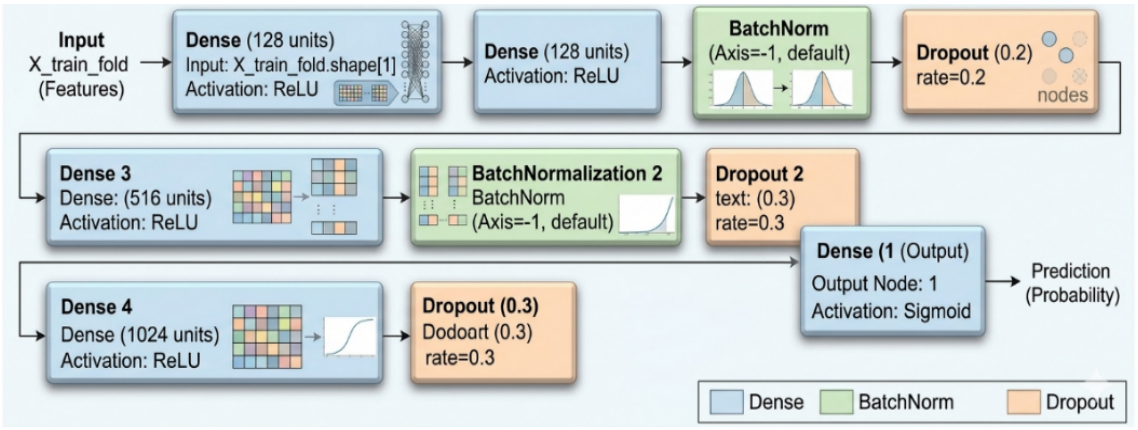

Figure 3. NN architecture design

The architecture of the NN model constructed using the Sequential Deep Learning architecture is depicted in Figure 3. It consists of four primary Dense layers with a neuron expansion scheme (128, 128, 516, and 1024 units). The ReLU activation function is used by each hidden layer to address the vanishing gradient and speed up convergence. The model incorporates regularization strategies like Dropout (rate 0.2–0.3) to avoid overfitting and Batch Normalization to stabilize the input distribution between layers in order to preserve performance and generalization. The output layer of this design consists of a single neuron unit that uses Sigmoid activation to provide binary classification probabilities between 0 and 1. The Neural Network model's training results are displayed in Table 3.

Table 3. Result of Training Model NN

Fold	Epochs	Validation Accuracy	Validation Loss	Average Validation Accuracy	Average Validation Loss	Standard Deviation Accuracy	Standard Deviation Loss
1	10	0.94	0.20				
2	10	0.94	0.16				
3	10	0.97	0.09	0.95	0.14	0.01	0.03
4	10	0.96	0.11				
5	10	0.96	0.15				

The NN model's assessment results utilizing a 5-fold cross-validation scheme are shown in Table 3, which demonstrates a very stable and dependable performance level. The model had a low average validation loss of 0.14 and an average validation accuracy of 0.95. Very low standard deviation values, 0.01 for accuracy and 0.03 for loss, support this stability and suggest that the model has excellent generalization capabilities and does not exhibit considerable bias toward certain data splits (folds). With an accuracy of 0.97 and the lowest loss of 0.09, the third fold exhibits peak performance, demonstrating that the employed architecture may reach an ideal convergence point in 10 epochs. Figure 4 and Figure 5 shows the accuracy and loss results during training.

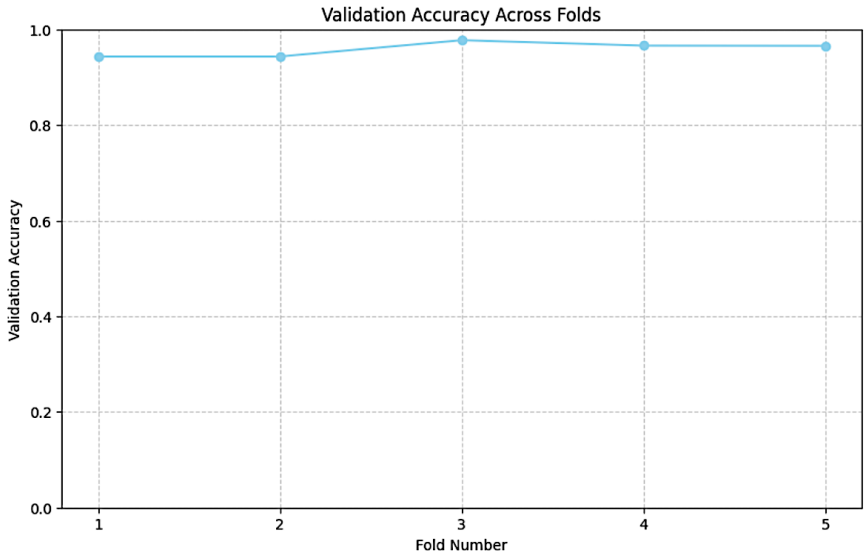

Figure 4. Results of Accuracy Across Folds Validation

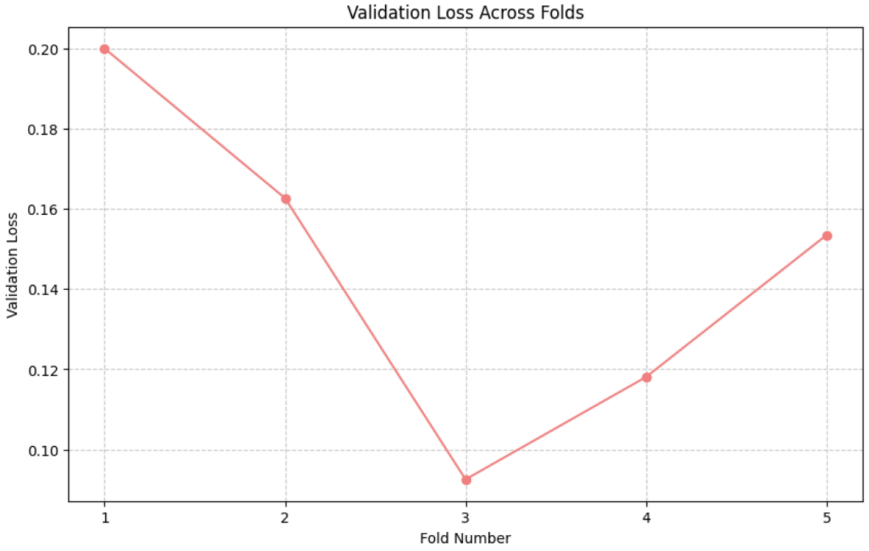

Figure 5. Results of Loss Across Folds Validation

The validation accuracy in Figure 4 exhibits a very steady and constantly high trend, staying over 0.95 throughout all folds, demonstrating the model's extremely robust and dependable predictive capabilities. Accordingly, Figure 5 displays variations in the validation loss value that often decline and stay in a narrow range (between 0.09 and 0.20), with the third fold serving as the ideal convergence point. When taken as a whole, this data demonstrates that the model can achieve great generalization with little variation across data subsets, which makes it ideal for use in the classification job. In the meantime, the SVM algorithm testing results are displayed in Table 4.

Table 4. Performance of the SVM algorithm

Model	Fold	Parameter	Accuracy for Each Fold	Mean Accuracy	Standard Deviation
SVM	1	linear; C=2;	0.95	0.96	0.01
	2	degree=3; scale	0.96		
	3		0.95		
	4		0.98		
	5		0.96		
	1	rbf; C=2; degree=4;	0.95	0.96	0.01
	2	scale	0.94		
	3		0.96		
	4		0.97		
	5		0.96		
	1	poly; C=5;	0.96	0.98	0.01
	2	degree=3; scale	0.98		
	3		0.96		
	4		0.98		
	5		1.0		

The SVM algorithm's experimental results with different kernel settings using a 5-fold cross-validation scheme are shown in Table 4. With an average accuracy of 0.98, the SVM model performs exceptionally well and steadily overall. The best results are obtained when the polynomial (poly) kernel with parameters is used. In fact, the model attained a perfect accuracy of 1.0 in the fifth fold of this arrangement. The extremely low standard deviation constantly stays at 0.01 across all tests, despite differences in the application of the linear and Radial Basis Function (RBF) kernels, which both yield an average accuracy of 0.96. This shows that, independent of the kind of kernel employed, the SVM algorithm handles this dataset with minimal variance and great dependability. These findings confirm that SVM, particularly when optimized with a polynomial kernel, is a very competitive model for this classification problem [16]. Meanwhile, Table 5 shows the results of testing the Random Forest algorithm.

Table 5. Performance of the RF algorithm

Model	Fold	Parameter	Accuracy for Each Fold	Mean Accuracy	Standard Deviation
RF	1	(max_depth=5, random_state=10, n_estimators=500, criterion='entropy', class_weight = None)	0.98	0.99	0.004
	2		1.0		
	3		1.0		
	4		1.0		
	5		1.0		

The assessment results of the RF algorithm set up with 500 estimators and the entropy criterion are shown in Table 5. With a mean accuracy of 0.99, the testing findings demonstrate that RF offers nearly flawless performance. Four folds that attained absolute accuracy (1.0) demonstrate this model's extremely high precision level [17], [18], and [19]. The assessment results of the RF algorithm set up with 500 estimators and the entropy criterion are shown in Table 5. With a mean accuracy of 0.99, the testing findings demonstrate that Random Forest offers nearly flawless performance. Four folds that attained absolute accuracy (1.0) demonstrate this model's extremely high precision level [17], [18], and [19]. The algorithm's exceptional stability against changes in the training data is statistically demonstrated by the extremely low standard deviation value of 0.004. By using the max_depth=5 option, the risk of overfitting is reduced because the model may achieve optimal generalization without requiring an excessive tree depth. In contrast, Table 6 presents the results of the NB algorithm.

Table 6. Performance of the NB algorithm

Model	Fold	Parameter	Accuracy for Each Fold	Mean Accuracy	Standard Deviation
NB	1	(class_prior=None, fit_prior=False, force_alpha=False, alpha=1.0)	0.93	0.94	0.03
	2		0.96		
	3		0.88		
	4		0.97		
	5		0.95		

The classification performance using the NB algorithm with an alpha parameter of 1.0 is shown in Table 6. This model obtained a competitive mean accuracy of 0.94 based on the 5-fold cross-validation test. However, compared to the prior model, the findings demonstrate a more dynamic accuracy variation, with the lowest value being 0.88 in the third fold and the highest value being 0.97 in the fourth. A modest degree of prediction variability across data subsets is shown statistically by a standard deviation value of 0.03. The model provides equal weights to each class without being affected by the data's initial frequency distribution when the fit_prior=False setting is used. While Naive Bayes performs well overall, it is not as stable as RF or SVM algorithms. Meanwhile, Table 7 shows the results of testing the kNN algorithm.

Table 7. Performance of the kNN algorithm

Model	Fold	Parameter	Accuracy for Each Fold	Mean Accuracy	Standard Deviation
kNN	1	(n_neighbors=5, weights='uniform', metric='minkowski', metric_params=None, n_jobs=None)	0.96	0.97	0.01
	2		0.98		
	3		0.98		
	4		0.98		
	5		0.95		

The kNN algorithm's performance evaluation results are displayed in Table 7. With an average accuracy (Mean Accuracy) of 0.97, the kNN model performs exceptionally well and consistently throughout. This approach is successful in mapping data patterns based on feature proximity since the accuracy on each fold is kept within a small range, between 0.95 and 0.98. The kNN model exhibits excellent stability against changes in the training data, as evidenced by the extremely low standard deviation value of 0.01. Each nearest neighbor is treated similarly when uniform weights are used, which has been shown to result in strong generalization in this instance.

3.6. Design Widget Orange

The Orange widget design [20] is shown in Figure 6, which is utilized to test the ROC analysis [21] on the accuracy measurement results of the OSCE exam passing status using the NB, kNN RF, SVM, and NN algorithms.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

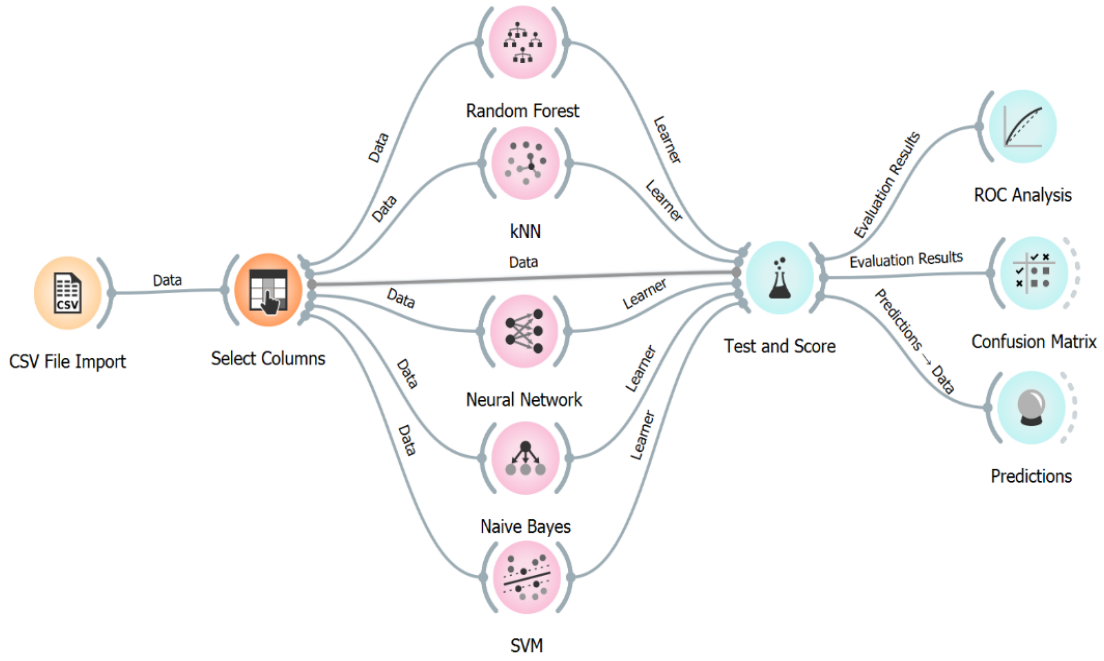

Figure 6. Design Widget Orange

The Orange Data Mining visual programming tool's data processing pipeline is depicted in Figure 6. From an academic perspective, this architecture shows how to integrate a structured classification pipeline, beginning with the pretreatment step of feature cleaning or selection (Select Columns) and data import (CSV File Import). Five distinct machine learning methods are then used in parallel to the chosen features: RF, kNN, NN, NB, and SVM. In order to compare each model's predicted performance fairly, the output from each model is combined in the Test and Score module. Lastly, our system generates performance evaluations in the form of confusion matrices, ROC curve analyses, and individual predictions, enabling the best model to be chosen using stringent evaluation measures.

3.7. Performance Evaluation

The Confusion Matrix, which assesses the model's prediction ability for the "Pass" and "Not Pass" classes, is shown in Figure 7. The model's performance in categorizing the test data samples with a total of 88 observations is displayed in the confusion matrix. Eight samples were correctly classified as "Not Pass" (True Negative) and 79 as "Pass" (True Positive) by the model; just one sample was incorrectly classified as "Pass" (False Positive). The "Pass" class had a flawless recall rate (100%) because no "Pass" samples were

incorrectly labeled as "Not Pass" (False Negative) [22]. Table 8 shows the evaluation results for the NN.

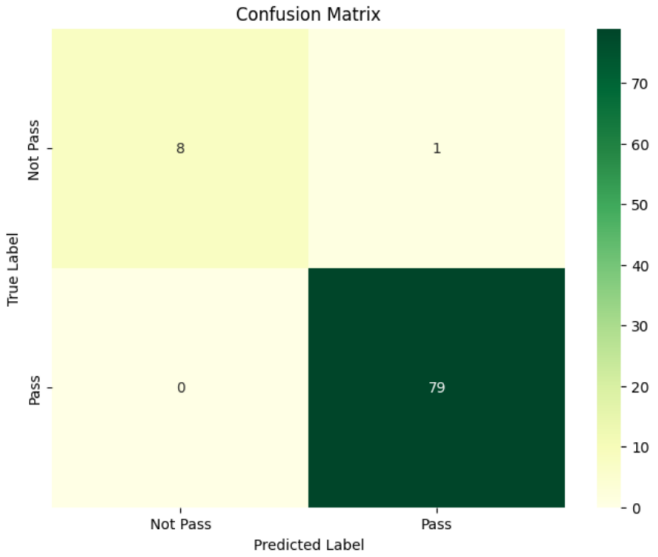

Figure 7. Confusion matrix NN algorithm

Table 8. NN algorithm's performance evaluation

	Precision	Recall	F1-Score	Class Data
Weighted Avg.	1.0	0.89	0.94	Not Pass
	0.99	1.0	0.99	Pass
	0.99	0.99	0.99	

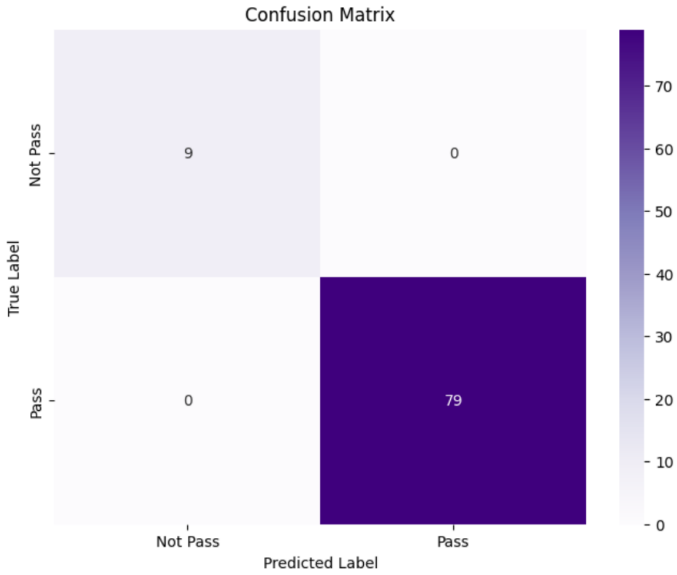

Figure 8. Confusion matrix SVM algorithm

The Confusion Matrix that assesses the model's prediction ability for the "Pass" and "Not Pass" classes is shown in Figure 8. The model properly categorized all 88 test samples with no prediction errors, according to the highly excellent findings displayed in the confusion matrix. In particular, there were 79 samples in the "Pass" category that were correctly predicted (True Positives) and 9 samples in the "Not Pass" category that were correctly predicted (True Negatives), with both False Positives and False Negatives being 0. Table 9 shows the evaluation results for the SVM.

Table 9. SVM algorithm's performance evaluation

	Precision	Recall	F1-Score	Class Data
	1.0	1.0	1.0	Not Pass
	1.0	1.0	1.0	Pass
Weighted Avg.	1.0	1.0	1.0	

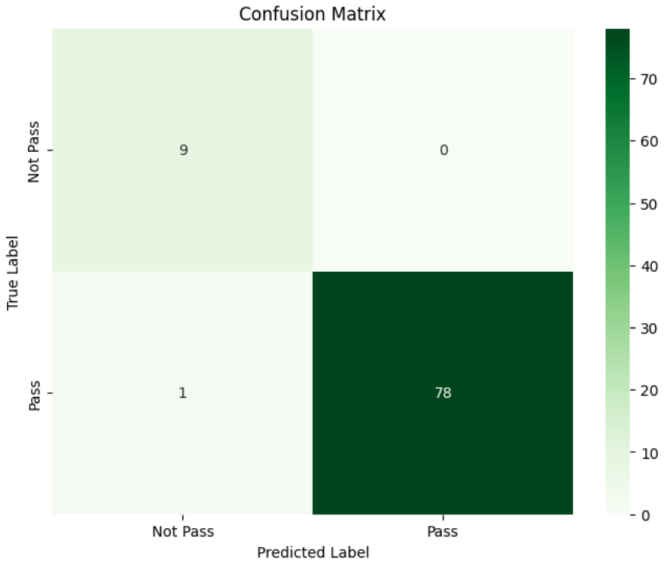

Figure 9. Confusion matrix RF algorithm

The Confusion Matrix that assesses the model's prediction ability for the "Pass" and "Not Pass" classes is shown in Figure 9. With a total of 88 test samples, the confusion matrix demonstrates a very strong performance, with the model accurately identifying 78 samples as "Pass" (True Positive) and 9 samples as "Not Pass" (True Negative). While there are no classification mistakes from "Not Pass" to "Pass" (False Positive), one "Pass" sample was mistakenly classified as "Not Pass" (False Negative). Tabel 10 shows the evaluation results for the RF.

Table 10. Results of the RF algorithm's performance evaluation

	Precision	Recall	F1-Score	Class Data
Weighted Avg.	0.90	1.0	0.95	Not Pass
	1.0	0.99	0.99	Pass
	0.99	0.99	0.99	

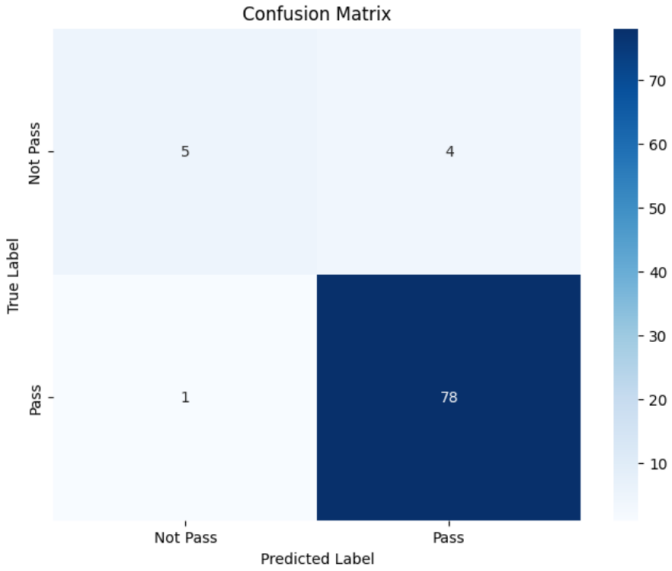

Figure 10. Confusion matrix NB algorithm

The Confusion Matrix, which assesses the model's prediction ability for the "Pass" and "Not Pass" classes, is shown in Figure 10. The confusion matrix successfully identifies 5 samples as "Not Pass" (True Negative) and 78 samples as "Pass" (True Positive), demonstrating high discriminative capabilities. One "Pass" sample was mistakenly classed as "Not Pass" (False Negative), and four "Not Pass" samples were mistakenly classified as "Pass" (False Positive). Table 11 shows the evaluation results for NB.

Table 11. Results of the NB algorithm's performance evaluation

	Precision	Recall	F1-Score	Class Data
Weighted Avg.	0.83	0.56	0.67	Not Pass
	0.95	0.99	0.97	Pass
	0.94	0.94	0.94	

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

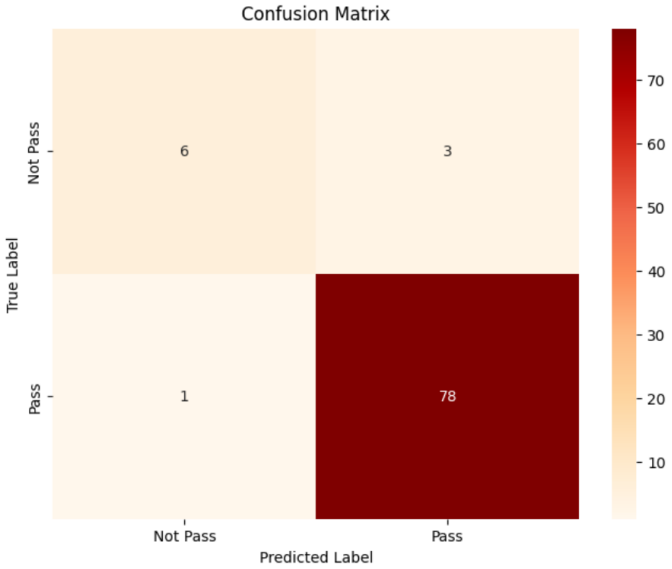

Figure 11. Confusion matrix kNN algorithm

The Confusion Matrix that assesses the model's prediction ability for the "Pass" and "Not Pass" classes is shown in Figure 11. The confusion matrix demonstrates good classification performance, correctly classifying 78 samples as "Pass" (True Positive) and 6 samples as "Not Pass" (True Negative). One "Pass" sample was mistakenly categorized as "Not Pass" (False Negative), whereas three samples from the "Not Pass" category were mistakenly forecasted as "Pass" (False Positive). Table 12 shows the evaluation results for kNN.

Table 12. Results of the kNN algorithm's performance evaluation

	Precision	Recall	F1-Score	Class Data
	0.86	0.67	0.75	Not Pass
	0.96	0.99	0.97	Pass
Weighted Avg.	0.95	0.95	0.95	

3.8. ROC Analysis

The performance of the NN, SVM, RF, NB, and kNN algorithms is displayed [23], managed, and categorized using Receiver Operating Characteristics (ROC). The NB, kNN RF, SVM, and NN algorithms' performance curves with the target classes "Pass" and "Not Pass," are displayed in Figure 12.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

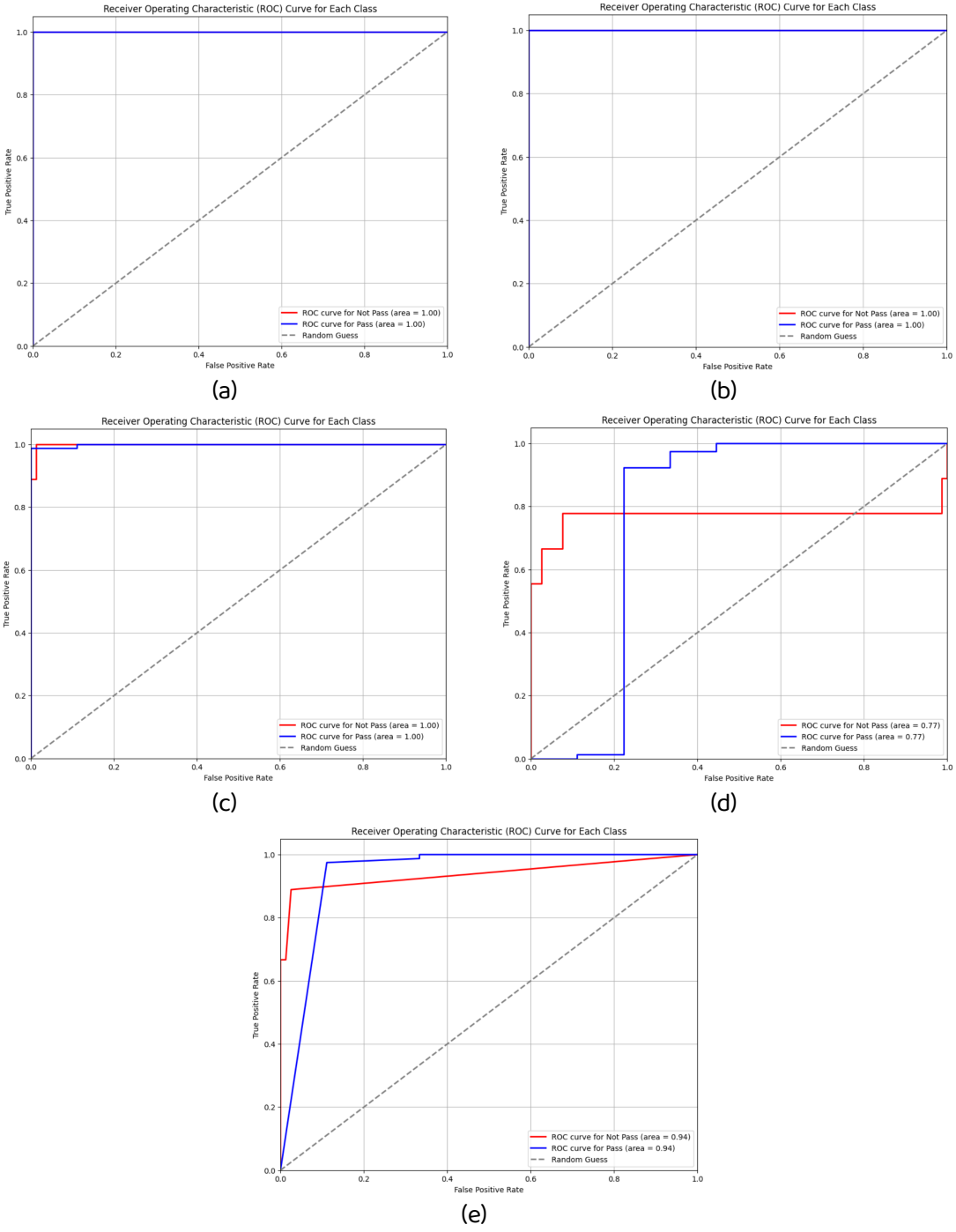

Figure 12. Performance ROC of the (a) NN, (b) SVM, (c) RF, (d) NB, and (e) kNN algorithms

The ROC curve illustrating the trade-off between True Positive Rate (sensitivity) and False Positive Rate (1-specificity) at different classification thresholds is displayed in Figure 12. With an AUC value of 1.00, which indicates faultless classification capacity without errors in differentiating between the "Pass" and "Not Pass" classes, the NN model

(a), SVM (b), and RF (c) exhibit exceptional classification performance based on the visualization. Conversely, NB (d) recorded an AUC of 0.77, showing poorer classification ability in comparison to other models, while kNN (e) model, with an AUC of 0.94, demonstrates very competitive performance. Overall, the integration of these ROC curves offers solid empirical support for the suggested models' dependability, with RF, SVM, and NN being the best methods for mapping intricate data features.

3.9. Comparison of NN, SVM, RF, NB, and kNN Algorithms against the Actual OSCE Exam Passing Status

Figure 13 shows the comparison of the NN, SVM, RF, NB, and kNN algorithms against the actual OSCE exam passing status, while Table 13 shows the validation results of the actual OSCE exam passing status compared to the OSCE exam passing status results using multiple algorithms.

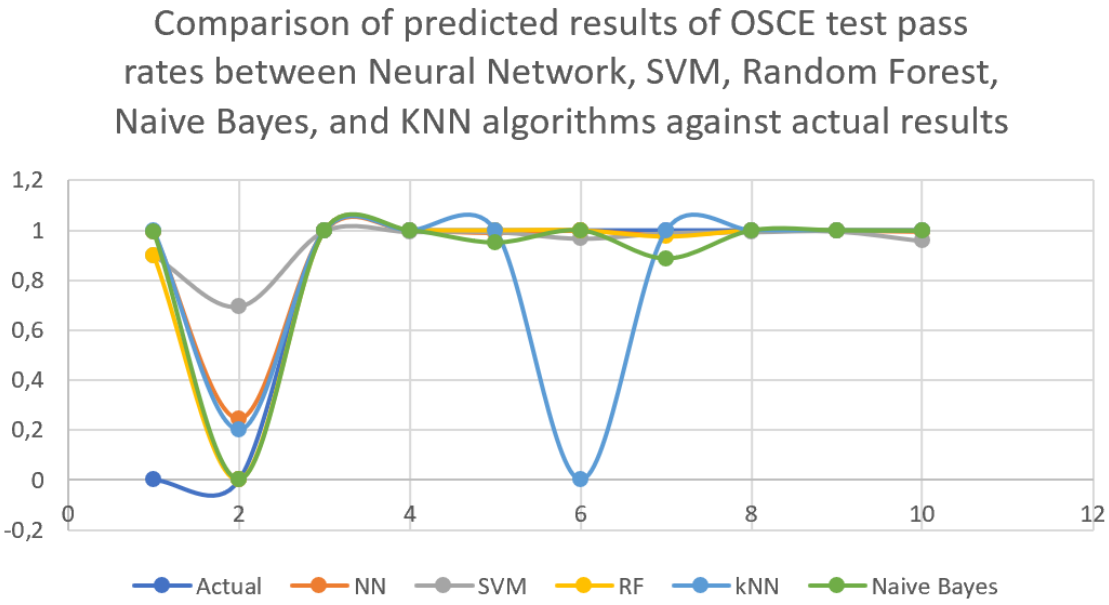

Figure 13. Performance curves of the NN, SVM, RF, NB, and kNN algorithms

The study is enhanced by Figure 13, which displays performance curves that illustrate how each model deviates from the real data throughout the course of the testing series. The overall goal of this visualization is to assess how well and consistently each machine learning model classifies graduation status depending on the input factors that are utilized. For ten student samples (MHS-01 to MHS-10), Table 13 displays numerical data on

prediction probabilities along with the ultimate status (pass/fail). The accuracy of each algorithm in reaching the real numbers varies.

Table 13. Validation results of the actual OSCE exam passing status compared to the OSCE exam passing status results using the NN, SVM, RF, NB, and kNN algorithms

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
MHS-01	0	1	1	1	1	1	0.9	Pass	0.8	Pass	0.9	Pass	0.9	Pass	1	Pass
MHS-02	0	1	1	1	1	1	0.2	Not Pass	0.6	Pass	0	Not Pass	0	Not Pass	0.2	Not Pass
MHS-03	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	1	Pass
MHS-04	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	1	Pass
MHS-05	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	1	Pass
MHS-06	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	0	Not Pass
MHS-07	1	1	1	1	1	1	0.9	Pass	0.9	Pass	0.9	Pass	0.8	Pass	1	Pass
MHS-08	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	1	Pass
MHS-09	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	1	Pass
MHS-010	1	1	1	1	1	1	0.9	Pass	0.9	Pass	1	Pass	0.9	Pass	1	Pass

Note: 1=Student Code, 2=Actual, 3=NN, 4=SVM, 5=RF, 6=NB, 7=kNN, 8= NN Numerical, 9= NN Validation Against Actual, 10=SVM Numerical, 11=SVM Validation Against Actual, 12=RF Numerical, 13=RF Validation Against Actual, 14=NB Numerical, 15=Validation of NB Against Actual, 16= kNN Numerical, 17= Validation of kNN Against Actual

3.10. Discussion

By filling a significant vacuum in previous research, this study expands the body of knowledge by presenting a multi-algorithm prediction framework for evaluating students' passing status in the Objective Structured Clinical Examination (OSCE). Predictive analytics for early identification of student performance results has rarely been examined in previous research, which has mostly concentrated on OSCE implementation, student perspectives, or digital delivery systems. Thus, this work's main contribution is to show how machine learning may be operationalized as a decision-support tool in competency-based medical education, in addition to attaining excellent predictive accuracy. All of the assessed algorithms NN, SVM, RF, kNN, and NB achieve accuracy levels above 90%, according to the experimental data. The chosen clinical competency indicators have significant predictive value and are highly relevant to OSCE outcomes, as evidenced by their continuously high performance. The machine learning

models used in this study effectively capture intricate, non-linear interactions among clinical skill factors, in contrast to conventional statistical methods that usually presume linear connections. The claim that OSCE performance is a multifaceted concept impacted by interconnected abilities rather than discrete skills is supported by this data.

Additionally, a study by [12] examines how students see AI-assisted OSCE preparation, placing greater emphasis on learning experience than predictive abilities. Similar to this, [13] used network analysis to comprehend the connections between OSCE components; however, their methodology is still descriptive and does not include predictive modeling. Together, these results show that while OSCE has been extensively studied, prediction frameworks that can facilitate early academic intervention are still receiving little attention.

On the other hand, by forecasting students' pass/fail status before the actual OSCE, this study presents a multi-algorithm machine learning framework that immediately closes this gap. This strategy is consistent with more general machine learning applications in healthcare, including the work of [7], which showed how well multi-algorithm systems predict illness status. However, the current study applies a similar methodological paradigm to educational evaluation, notably focusing on competency-based outcomes, in contrast to earlier studies that concentrate on clinical diagnosis. By connecting machine learning techniques with medical education assessment systems, this makes a unique contribution.

Although technically possible, the SVM model's 100% accuracy needs to be carefully interpreted in light of the dataset and experimental setup. When the feature space is substantially separable that is, when the chosen clinical competency factors create distinct decision boundaries between the "Pass" and "Not Pass" classes such flawless classification performance can take place. This separability was probably made possible in this work by the inclusion of normalized, structured, and domain-relevant characteristics, which allowed SVM, especially with a polynomial kernel, to create an ideal hyperplane with maximum margin. However, a number of risk factors need to be taken into account. First, the model is more likely to capture dataset specific patterns rather than generalizable associations due to the relatively small dataset size (n = 439), which could result in overfitting. Second, the use of cross validation without external validation

may still allow subtle data leakage or distributional similarity between folds, artificially inflating performance metrics. Third, the high dimensional consistency and possible correlation among features may reduce variability in the data, further simplifying the classification task. Additionally, class balance (221 pass vs. 218 not pass) may contribute to stable training but does not eliminate the risk of over-optimistic evaluation. Therefore, while the 100% SVM result indicates strong model capability, it should be interpreted as an upper-bound performance under controlled conditions rather than definitive evidence of real-world generalization. Future validation using larger, multi-institutional datasets and stricter evaluation protocols such as hold-out testing or external validation are necessary to confirm the robustness and practical applicability of this finding. The Random Forest model likewise showed near optimal performance. This implies that because ensemble learning techniques successfully balance bias and variance while preserving high predicted accuracy, they are especially well suited for this classification task. When it comes to model interpretability, Random Forest is superior to SVM, especially when it comes to feature importance analysis. This ability is extremely important in educational settings because it allows teachers to determine which clinical competencies have the biggest impact on students' success or failure.

The NN model demonstrated its capacity to represent intricate patterns in the dataset with a robust average accuracy of 95%. The low standard deviation across cross-validation folds indicates that the use of batch normalization and dropout led to consistent training and decreased overfitting. However, the NN model requires more intense computational resources and parameter adjustment than SVM and RF, which may restrict its scalability in real-world educational settings with constrained technical infrastructure. Additionally, the kNN algorithm yielded competitive results (97% accuracy), demonstrating that instance-based learning techniques continue to be successful when the feature space is normalized and well-structured. However, it is computationally less effective for larger datasets and susceptible to feature scaling because to its reliance on distance measures. Naive Bayes, on the other hand, performed the worst out of all the models that were assessed, mostly because of its feature independence assumption, which is not entirely consistent with the intrinsically coupled nature of clinical competency variables.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Instead of being based on a single hold-out test set, the evaluation metrics presented in this work, such as the confusion matrix and ROC analysis, are obtained from the 5-fold cross-validation technique and represent aggregated findings across all folds. In particular, every observation was utilized exactly once as validation data because the model was trained on 80% of the data in each fold and validated on the remaining 20%. The model's overall classification performance on unknown data is reflected in the confusion matrix values (True Positive, True Negative, False Positive, and False Negative) displayed in the results, which match the total predictions from all validation folds. In a similar vein, the aggregated probability outputs from all folds were used to calculate the ROC curves and AUC values, yielding a more consistent and trustworthy indication of the model's capacity for discrimination. In order to minimize variance in performance estimation and make the most use of the available dataset, this cross-validated aggregation approach was selected. It should be highlighted, nonetheless, that these findings do not represent a true hold-out or external validation. Therefore, future research should include an independent test set or external dataset to further assess the models' generalizability, even though the given metrics offer compelling evidence of internal model consistency.

This study contains a number of shortcomings that should be addressed in subsequent research, despite its merits. First, the results may not be as broadly applicable due to the dataset's small size and single institution origin. Second, other potentially significant aspects like cognitive capacity, communication skills, psychological preparation, or learning behaviors are not taken into account by the model; instead, it just takes into account quantitative clinical skill variables. Third, while 5-fold cross-validation was used for internal validation, independent datasets were not used for external validation, which is crucial for verifying the robustness of the model. Future studies should concentrate on growing the dataset across several institutions and adding more varied variables, such as longitudinal and behavioral data. Furthermore, it is advised to use explainable AI (XAI) methodologies to enhance model transparency and facilitate interpretability for decision-makers and educators. Predictive performance may be further improved by investigating hybrid or ensemble models that incorporate the advantages of several techniques. This study concludes by showing that predictive modeling based on machine learning is a viable strategy for enhancing OSCE assessment systems. Educational institutions can better promote student achievement and guarantee greater competency

standards in healthcare education by transitioning from reactive assessment approaches to proactive, data-driven decision-making.

4. CONCLUSION

This study compared the performance of five machine learning algorithms SVM, RF, kNN, NN, and NB in predicting OSCE passing status using clinical competency data. All models achieved accuracy above 90%, with SVM and RF consistently showing the best performance in terms of accuracy and stability, followed by NN and kNN, while NB performed relatively lower. These results highlight the importance of comparative evaluation in identifying the most suitable algorithm for structured educational data. However, the findings are limited by the use of a single-institution dataset and the absence of external validation. Therefore, broader validation is required to ensure model generalizability across different educational settings. Future work should focus on expanding the dataset across multiple institutions, incorporating additional variables (e.g., behavioral and cognitive factors), and applying external validation or independent test sets. Further exploration of ensemble or explainable AI approaches is also recommended to improve both performance and interpretability.

ACKNOWLEDGMENT

The author would like to express their sincere gratitude to Aisyah University for the support and facilities provided during the completion of this research.

REFERENCES

[1] A. A. Elbilgahy, F. Eltaib, F. A. Eltaib, and R. K. KMohamed, "Implementation of Objective Structured Clinical Examination (OSCE): Perceiving Nursing Students and Teachers Attitude & Satisfaction," *Am. J. Nurs. Res.*, vol. 8, no. 2, pp. 220–226, 2020, doi: 10.12691/ajnr-8-2-11.

[2] Z. Zulkifli, R. Ratnasari, Y. Arifin, and C. Habib, "Agile-Scrum Methodology for Hospital Information System Development," *J. Inf. Syst. Informatics*, vol. 7, no. 2, pp. 1696–1713, 2025, doi: 10.51519/journalisi.v7i2.1148.

[3] Z. Zulkifli, M. Mardiana, and D. Despa, "Software Quality Assessment Model: A New

Approach for Software Testing Tools," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 35, no. 2, pp. 139–155, 2024, doi: 10.1142/S0218194024500517.

[4] S. Lebdai, B. Bouvard, L. Martin, C. Annweiler, N. Lerolle, and E. Rineau, "Objective structured clinical examination versus traditional written examinations: a prospective observational study," *BMC Med. Educ.*, vol. 23, no. 1, pp. 1–9, 2023, doi: 10.1186/s12909-023-04050-5.

[5] Z. Zulkifli, S. Wahyuningsih, A. I. Hendrawan Putri, T. H. Andika, P. Bintoro, and M. Pratiwi, "Implementation of a Diabetes Status Prediction Application Using a Machine Learning Algorithm Approach," *Proc. 2024 Int. Conf. Inf. Manag. Technol. ICIMTech 2024*, pp. 701–706, 2024, doi: 10.1109/ICIMTech63123.2024.10780853.

[6] Zulkifli, F. L. Gaol, A. Trisetyarso, and W. Budiharto, "Software Testing Model by Measuring the Level of Accuracy Fault Output Using Neural Network Algorithm," *Proc. 2022 IEEE 7th Int. Conf. Inf. Technol. Digit. Appl. ICITDA 2022*, pp. 1–6, 2022, doi: 10.1109/ICITDA55840.2022.9971274.

[7] Z. Zulkifli, F. A. Makkiyah, D. Antoni, F. Fitriana, T. Jamaan, and A. Taufik, "Multi-Algorithm to Measure the Accuracy Level of Diabetes Status Prediction," *J. Appl. Data Sci.*, vol. 5, no. 2, pp. 736–746, 2024, doi: 10.47738/jads.v5i2.250.

[8] T. A. Hannan, S. Y. Umar, Z. Rob, and R. R. Choudhury, "Designing and running an online Objective Structured Clinical Examination (OSCE) on Zoom: A peer-led example," *Med. Teach.*, vol. 43, no. 6, pp. 651–655, 2021, doi: 10.1080/0142159X.2021.1887836.

[9] J. F. Naga and R. Q. Lavilles, "Deciphering Digital Discourse: Detecting Cyberbullying Patterns in Filipino Tweets Using Machine Learning," *CommIT J.*, vol. 18, no. 2, pp. 167–181, 2024, doi: 10.21512/commit.v18i2.11094.

[10] T. K. Soong and C. M. Ho, "Artificial intelligence in medical OSCEs: Reflections and future developments," *Adv. Med. Educ. Pract.*, vol. 12, pp. 167–173, 2021, doi: 10.2147/AMEP.S287926.

[11] M. Tekin, M. O. Yurdal, Ç. Toraman, G. Korkmaz, and İ. Uysal, "Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination," *BMC Med. Educ.*, vol. 25, no. 1, 2025, doi: 10.1186/s12909-025-07241-4.

[12] S. Rehman, M. Ali, E. Cheema, and A. Shanzeh, "Exploring the perceptions and experiences of pharmacy students about formative and summative OSCE incorporating AI in preparatory process: A mixed-methods study," *Curr. Pharm.*

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Teach. Learn., vol. 17, no. 6, p. 102348, 2025, doi: https://doi.org/10.1016/j.cptl.2025.102348.

[13] H. Zhang *et al.*, "Network analysis of an OSCE-based graduation skills assessment for clinical medical students," *BMC Med. Educ.*, vol. 25, no. 1, 2025, doi: 10.1186/s12909-025-07091-0.

[14] W. Chine, "A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks," *Renew. Energy*, vol. 90, pp. 501–512, 2016, doi: 10.1016/j.renene.2016.01.036.

[15] H. Wang, R. Czerminski, and A. C. Jamieson, "Neural Networks and Deep Learning," *Mach. Age Cust. Insight*, pp. 91–101, 2021, doi: 10.1108/978-1-83909-694-520211010.

[16] P. Brar and D. Nandal, "A Systematic Literature Review of Machine Learning Techniques for Software Effort Estimation Models," *Proc. - 2022 5th Int. Conf. Comput. Intell. Commun. Technol. CCICT 2022*, pp. 494–499, 2022, doi: 10.1109/CCiCT56684.2022.00093.

[17] D. R. Ibrahim, "Software defect prediction using feature selection and random forest algorithm," 2017. doi: 10.1109/ICTCS.2017.39.

[18] S. Liu, "An Integrated Scheme for Online Dynamic Security Assessment Based on Partial Mutual Information and Iterated Random Forest," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3606–3619, 2020, doi: 10.1109/TSG.2020.2991335.

[19] C. C. By-nc-nd, "Zupan, Demsar: Introduction to Data Mining," no. May, pp. 1–72, 2018.

[20] Z. Zulkifli, F. L. Gaol, T. Agung, and B. Widodo, "Software Testing Integration-Based Model (I-BM) Framework for Recognizing Measure Fault Output Accuracy Using Machine Learning Approach," *Int. J. Softw. Eng. Knowl. Eng.*, 2023.

[21] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[22] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, vol. 148, pp. 56–62, 2018, doi: 10.1016/j.beproc.2018.01.004.

[23] C. Catal, O. Alan, and K. Balkan, "Class noise detection based on software metrics and ROC curves," *Inf. Sci. (Ny).*, vol. 181, no. 21, pp. 4867–4877, 2011, doi: 10.1016/j.ins.2011.06.017.