IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2484~2493
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2484-2493

Prediction of new student admissions to higher education using support vector machines

Neni Purwati[1], Windya Harieska Pramujati[2], A. Aviv Mahmudi[3], Mira Febriana Sesunan[4], Yahya[4]
[1]Department of Medical Informatics, Faculty of Health Science, Universitas Muhammadiyah Lamongan, Lamongan, Indonesia
[2]Department of Information Technology, Politeknik Negeri Malang, Malang, Indonesia
[3]Department of Information Systems, Faculty of Science and Technology, Universitas YPPI Rembang, Rembang, Indonesia
[4]Department of Information Systems, Faculty of Engineering, Universitas Darma Persada, Jakarta, Indonesia

Article Info

Article history:

Received Aug 6, 2024
Revised Jan 13, 2026
Accepted May 11, 2026

Keywords:

Higher education
Kernel
New student admissions
Prediction
Support vector machine

ABSTRACT

Higher education institutions across various regions operate using systems that generate large amounts of data. This data is stored and utilized for strategic decision-making, providing significant business value to these institutions. Support vector machine (SVM) has become popular due to its strong generalization capability, high prediction accuracy, and faster training speed. SVM employs kernels as tuning parameters. This study aims to enhance the accuracy of student admissions prediction in higher education institutions using the SVM classification model. The SVM model was applied to a dataset comprising 5,936 records with four attributes and was evaluated using the use training set, 10-fold cross-validation, and percentage splits of 70%–30% and 80%–20%. Initially, the SVM-kernel model achieved high accuracy but failed to identify any true positive instances, indicating its inability to detect the minority "not accepted" class due to severe class imbalance. After applying class balancing techniques, the model's performance improved significantly in terms of area under the curve (AUC), F-measure, and Matthews correlation coefficient (MCC), reflecting a more balanced classification between majority and minority classes. The SVM with Pearson VII function-based universal kernel (PUK) and classifier version 4.5 (C4.5) models achieved the best performance, indicating that class balancing effectively enhances both sensitivity and fairness in predictive classification.

Corresponding Author:

Neni Purwati
Department of Medical Informatics, Faculty of Health Science, Universitas Muhammadiyah Lamongan
St. Raya Plalangan No.KM 02, Lamongan, East Java, Indonesia
Email: nenipurwati@umla.ac.id

1. INTRODUCTION

In the 21st century, higher education has become a fundamental pillar of social and economic life. Additionally, the educational process ensures two other crucial aspects: a promising career and financial security. However, predicting student admissions to higher education institutions is challenging, as prospective students often lack knowledge of the admission requirements [1]. The student admission process is a crucial step from both the students' and the higher education institutions' perspectives [2]. Higher education institutions operate across various geographical regions using systems that generate large amounts of data. This data is stored and utilized by decision-makers to make strategic decisions, which significantly affect the business value of the institutions. Therefore, it can enhance the effectiveness of managerial decision-making [3]. Thus, this large amount of data can be optimally utilized to gain new insight. Machine

learning (ML), a subdomain of artificial intelligence (AI), has proven its performance across various fields. This study aims to enhance the accuracy of predicting student admissions to higher education institutions. Predicting new student admissions plays a crucial role in improving the efficiency, fairness, and overall quality of selection process in higher education. Such systems streamline evaluation procedures, minimize human bias, and enable universities to manage resources more effectively while enhancing academic standards. In addition, the prediction outcomes provide valuable input for strategic planning and data-driven decision-making in the administration of student admissions [4]. Previous studies have demonstrated that ML techniques are effective in enhancing prediction accuracy [5]. ML and AI technologies have significantly transformed society and various industrial sectors by automating critical decision-making processes, exerting a substantial impact on domains such as criminal law, healthcare, finance, and the workforce [6].

Among various ML approaches, support vector machines (SVMs) are distinguished for their ability to handle nonlinear patterns and high-dimensional datasets, characteristics that are frequently encountered in polymer research [7]. An effective classification approach using SVMs involves employing suitable feature selection methods that prioritize influential attributes while excluding less relevant ones. This scenario has the potential to enhance classification accuracy and reduce computational overhead. Furthermore, utilizing reliable datasets and applying appropriate validation techniques contributes to producing accurate and trustworthy results [8]. Wind speed prediction, for instance, is crucial for optimizing power generation and maintaining electricity supply stability, which utilized meteorological data from the National Wind Technology Center (NWTC) in Boulder, Colorado, employed three prediction models: fine tree, SVM, and linear regression, and demonstrated the importance of model evaluation in achieving reliable forecasts [9]. Similarly, an SVM model combined with stratified and shuffle sampling techniques demonstrated optimal and reliable performance in predicting students' academic outcomes [10]. Another study applied supervised learning algorithms, including SVM, decision tree (DT), random forest (RF), and extreme gradient boosting (XGB), to predict student academic achievement, with XGB achieving the highest F1-score and a success rate of 77% [11]. In addition, a web-based system employing RF, SVM, and DT algorithms was developed to predict student–course suitability, with RF achieving the highest accuracy of 95%, thereby enhancing academic guidance and reducing curriculum mismatches [12].

This section also reviews of several previous studies focusing on the application of ML methods in student admission prediction systems. The review aims to examine the approaches, algorithms, and findings of these studies, as well as to identify remaining research gaps. Table 1 provides a comparison of selected relevant studies that serve as the foundation for this research, particularly concerning the use of kernel-based SVM methods to enhance prediction accuracy. Referring to the analysis presented in Table 1 [13]–[17] (see in appendix), it is evident that further investigation is required to assess the effectiveness of kernel-based approaches. Therefore, this study evaluates the impact of various SVM kernels, including normalized polynomial, linear, radial basis function (RBF), and Pearson VII function-based universal kernel (PUK), on prediction accuracy and generalization capability using multiple evaluation metrics such as accuracy, recall, precision, F1-measure, Matthews correlation coefficient (MCC), and receiver operating characteristic (ROC). Moreover, this study emphasizes the importance of kernel selection as a critical factor in achieving accurate classification.

2. METHOD

2.1. Dataset

The case study was conducted at a university in Indonesia. The dataset consists of new student admissions records collected over a five-year period from 2016 to 2020, comprising 5936 records and 11 attributes. These attributes include number, name, major, year, gender, place of birth, date of birth, sub-district, district, age, and class. Not all features are incorporated into the predictive model, as their inclusion was determined based on relevance and redundancy criteria. Date of birth was excluded as age provides equivalent information, while number and name were removed as non-informative identifiers. Attributes such as gender, place of birth, sub-district, and district are omitted due to potential bias or limited generalizability, except in contexts where location-based or quota-driven admission policies apply. Conversely, major, year, and age were retained as key predictive variables, with class serving as the target variable. In total, four attributes were utilized in the analysis. A total of 23 missing values were identified in the sub-district and district attributes; however, these were excluded from the analysis due to their irrelevance. No duplicate records were detected after applying a duplicate-removal check.

2.2. Support vector machine

The SVM is one of the most widely used supervised ML algorithms for solving problems related to classification, regression, recognition, and time series analysis. SVM has gained significant popularity due to its features such as strong generalization ability, high prediction accuracy, and rapid training speed.

SVM performs well with small sample data sizes and exhibits high effectiveness in classifying both linear and non-linear data. In a binary classification problem, let there be several training samples represented as $X_i = \{X_1, X_2, X_3, \ldots, X_n\}$ in the input space where $X_i \epsilon Rd$ and their corresponding labels are represented as $y_i \epsilon \{-1, +1\}$ [18].

2.2.1. Support vector machine kernel function

Training samples that are located closer to the hyperplane are referred to as support vectors [19]. The maximum margin hyperplane is also known as the optimal separating hyperplane (OSH). Mapping training samples from the input space to higher-dimensional spaces is accomplished using a kernel function $k(x_i, x_j)$ as illustrated in Figure 1.

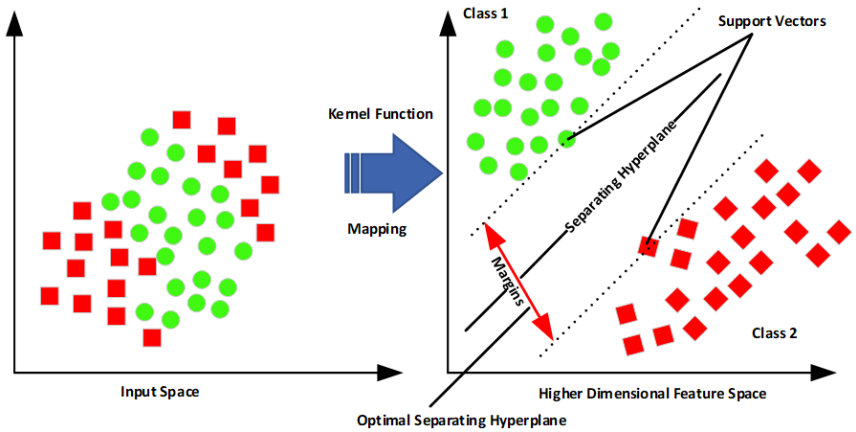

Figure 1. Kernel-based high-dimensional feature mapping [20]

SVM has kernels as tuning parameters. The kernels can be linear, RBF, sigmoid, or polynomial [21]. The proposed method is verified and compared with the RBF kernel, sigmoid kernel, and normalized poly kernel (NPK) SVM [22]. The NPK achieves similarity not only in the functions stated in the input samples but also in their combinations [23]. The linear kernel (LK) is effective when data can be separated linearly, while the NPK resembles the LK. In the Waikato environment for knowledge analysis (WEKA) data mining application, SVM is referred to as sequential minimal optimization (SMO) and consists of 6 kernels: NPK, poly/ LK, matrix kernel, PUK kernel, RBF kernel, and string kernel. In this study, only four kernel functions are used: NPK, LK, RBF kernel, and PUK kernel. The sigmoid kernel is not available in the latest version of WEKA because it does not always satisfy the positive semi-definite property, which often leads to numerical instability and convergence issues during SVM training, its performance also tends to be less consistent compared to the RBF or PUK kernel, therefore the WEKA developers decided to remove the sigmoid kernel and replace it with the PUK kernel, which offers greater stability, flexibility, and accuracy in data classification tasks. Formulas for the four kernel functions mentioned as in (1) to (4) [24].

$$Normalized\ Poly\ Kernel: K(x,y) = \frac{<x,y>^{2.0}}{(<x,x>^{2.0} * <y,y>^{2.0})^{(\frac{1}{2})}} \quad (1)$$

$$Poly\ Kernel: K(x,y) = <x,y>^{p} \quad (2)$$

$$RBF: K(x,y) = exp(-gamma * (x-y)^2) \quad (3)$$

$$PUK\ Kernel: K(x,x') = \frac{1}{(1+(\sqrt[2]{||x-x'||^2 \cdot 2^{\frac{1}{\omega}}-1})^2)^{\omega}} \quad (4)$$

2.3. Performance evaluation metrics

The classification performance is evaluated using performance measures such as accuracy, precision, recall, F-measure, and a confusion matrix. These metrics help assess classification errors, false positives (FP), and false negatives (FN). They provide a clear basis for measuring the model's classification quality.

i) Accuracy: accuracy is calculated using (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Where true positives (TP) refer to positives that are classified correctly, true negatives (TN) refer to negatives that are classified correctly, FP refer to negatives that are classified incorrectly, and FN refer to positives that are classified incorrectly [25], [26].

ii) Precision: precision is the proportion of correctly classified positive class instances out of the total classified instances of that class. Precision is measured using (6) [20].

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

iii) Recall: recall is a measure of testing accuracy considering precision and computational requirements, calculated using (7) [27].

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

iv) F1-measure: F1-measure is the mean between precision and recall. It ranges between 0 and 1. F1-measure is measured using (8) [20].

$$F1 - measure = 2 \times \frac{Precision \times ecall}{Precision+Recall} \quad (8)$$

v) MCC: MCC is crucial as it represents the correlation between the target and predicted values generated from classification, and is formulated as in (9) [28].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (9)$$

vi) ROC: the ROC curve represents the relationship between the FP and the TP for various threshold values, statistically used to evaluate the effectiveness of different classifiers. The area under the curve (AUC), where a classifier with an AUC of 1.0 indicates a high-level AUC and demonstrates better performance [29].

3. RESULTS AND DISCUSSION

This study employs data modelling using logistic regression (LR), Classifier version 4.5 (C4.5) DT, and naive Bayes (NB) as baseline classification models. The resulting accuracy values are presented in Table 2. The test results presented in Table 2 indicate that the three models: LR, C4.5, and NB, have relatively consistent performance across various validation methods. In both the use training set and cross validation fold-10 (CVF-10) tests, the accuracy of LR and C4.5 is nearly identical (approximately 95.7%), slightly higher than that of NB (94.9%). When tested with a 70% training and 30% testing data split, the C4.5 model obtained the highest accuracy of 96.23%, followed by LR with 96.18% and NB with 95.33%. The minimal variance in accuracy across models suggests strong generalization capability, although C4.5 performs marginally better due to its ability to capture non-linear attribute interactions and execute more effective data segmentation. In the evaluation with an 80% training and 20% testing data split, all models demonstrated improved accuracy. The C4.5 model maintained the highest performance (96.54%), followed by LR at 96.46% and NB at 95.45%. This enhancement suggests that increasing the proportion of training data enables the models to better capture underlying patterns, thereby enhancing predictive performance. These findings further confirm the consistent superiority of the C4.5 model among the three baseline algorithms.

Table 2. Model comparison

Test option/method	LR (%)	C4.5 (%)	NB (%)
Use training set (UTS)	95.78	95.75	94.91
CVF-10	95.77	95.75	94.91
Percentage split (PS)-70%	96.18	96.23	95.33
PS-80%	96.46	96.54	95.45

The research process was carried out by setting the hyperparameter c =1.0 for all kernels, gamma =0.01 for the RBF kernel, omega =1.0, and sigma =1.0 for the PUK kernel. The resulting accuracy

values for the four kernel types in the SVM model, based on the dataset described earlier, are summarized in Table 3. In Table 3, the test results indicate that all four SVM kernel types, normalized polynomial, linear, RBF, and PUK, achieved identical accuracy across all evaluation scenarios, including use training set, 10-fold cross validation, PS 70%, and 80%. The consistent accuracy values (ranging from 95.7% to 96.5%) indicate that the choice of kernel has no significant impact on the model's performance for this dataset. This indicates that the data patterns are relatively linear and not highly complex, allowing all kernels to separate the classes equally effectively. Furthermore, the accuracy improvement in the 70% and 80% PS scenarios demonstrates that increasing the amount of training data enhances the model's ability to recognize data patterns and improve predictive performance.

Table 3. Accuracy value results

Test option	Accuracy	Accuracy kernel type							
		Number of classes	NPK (%)	Number of classes	LK (%)	Number of classes	RBF (%)	Number of classes	PUK (%)
UTS	CCI	5,684	95.7	5,684	95.7	5,684	95.7	5,684	95.7
	ICI	252	4.2	252	4.2	252	4.2	252	4.2
CVF-10	CCI	5,684	95.7	5,684	95.7	5,684	95.7	5,684	95.7
	ICI	252	4.2	252	4.2	252	4.2	252	4.2
PS-70%	CCI	1,714	96.2	1,714	96.2	1,714	96.2	1,714	96.2
	ICI	67	3.7	67	3.7	67	3.7	67	3.7
PS-80%	CCI	1,146	96.5	1,146	96.5	1,146	96.5	1,146	96.5
	ICI	41	3.4	41	3.4	41	3.4	41	3.4

Description: CCI = correctly classified instances, ICC = incorrectly classified instances

Performance measure values consisting of precision, recall, and F-measure were also computed, and the results are summarized in Table 4. Table 4 shows that among the four kernel types and four evaluation methods (UTS, CVF-10, PS-70%, and PS-80%), the highest performance metrics were obtained using the PS-70%, for all kernels under this configuration, the precision, F-measure, and MCC values were undefined due to the absence of instances or predictions that enabled calculation for certain classes. However, the recall value reached 0.962, and the AUC was 0.500. When tested with PS-80%, the results indicated a slight improvement in recall for all kernels (0.965), while the other performance metrics remained unchanged.

Table 4. Performance measure weighted average value

Kernel type	Test option	AUC	Precision	Recall	F-measure	MCC
NPK	UTS	0.500	?	0.958	?	?
	CVF-10	0.500	?	0.958	?	?
	PS-70%	0.500	?	0.962	?	?
	PS-80%	0.500	?	0.965	?	?
LK	UTS	0.500	?	0.958	?	?
	CVF-10	0.500	?	0.958	?	?
	PS-70%	0.500	?	0.962	?	?
	PS-80%	0.500	?	0.965	?	?
RBF	UTS	0.500	?	0.958	?	?
	CVF-10	0.500	?	0.958	?	?
	PS-70%	0.500	?	0.962	?	?
	PS-80%	0.500	?	0.965	?	?
PUK	UTS	0.500	?	0.958	?	?
	CVF-10	0.500	?	0.958	?	?
	PS-70%	0.500	?	0.962	?	?
	PS-80%	0.500	?	0.965	?	?

The confusion matrix values from the conducted testing can be seen in Table 5. This table explains that the confusion matrix indicates that all four SVM kernels (normalized polynomial, linear, RBF, and PUK) yield a high number of TN while reporting zero TP, suggesting that the model classifies all instances exclusively as the negative class. Despite the seemingly high accuracy, the model is unable to identify the positive class, revealing a significant issue of class imbalance. The application of ClassBalancer serves as an efficient approach to address class imbalance without modifying the dataset by adding or removing instances, this filter functions by reweighting each instance, ensuring that the minority class exerts influence comparable to the majority class during model training. As a result, the model reduces its bias toward the dominant "accepted" class and becomes more responsive to the "unaccepted" class, thereby enhancing the

equilibrium between overall accuracy and minority class detection performance. The parameter c plays a crucial role in controlling the complexity of the SVM model by balancing margin maximization and tolerance to classification errors. An excessively large or small c value may result in overfitting or underfitting; therefore, determining an optimal c value is essential for maintaining model stability and accuracy, even when other kernel parameters (such as gamma, omega, and sigma) are properly configured. Following the application of the ClassBalancer technique, using hyperparameters c =1.0 for all kernels, gamma =0.01 for the RBF kernel, and omega =1.0 with sigma =1.0 for the PUK kernel, the achieved accuracy and corresponding confusion matrices are summarized in Table 6.

Table 5. Confusion matrix

Kernel type	Test option	AUC	Precision	Recall	F-measure	MCC
NPK	UTS	5684	0	252	0	UTS
	CVF-10	5684	0	252	0	CVF-10
	PS-70%	1714	0	67	0	PS-70%
	PS-80%	1146	0	41	0	PS-80%
LK	UTS	5684	0	252	0	UTS
	CVF-10	5684	0	252	0	CVF-10
	PS-70%	1714	0	67	0	PS-70%
	PS-80%	1146	0	41	0	PS-80%
RBF	UTS	5684	0	252	0	UTS
	CVF-10	5684	0	252	0	CVF-10
	PS-70%	1714	0	67	0	PS-70%
	PS-80%	1146	0	41	0	PS-80%
PUK	UTS	5684	0	252	0	UTS
	CVF-10	5684	0	252	0	CVF-10
	PS-70%	1714	0	67	0	PS-70%
	PS-80%	1146	0	41	0	PS-80%

Table 6. Confusion matrix and performance measure (balanced data)

Model	Kernel types	Test option	TN	TP	FN	FP	Accuracy	AUC	Precision	Recall	F-measure	MCC
SVM	NPK	UTS	1965.44	2920.89	47.11	1002.56	82.31	0.823	0.861	0.823	0.818	0.683
		CVF-10	1961.78	2920.89	47.11	1006.22	82.25	0.823	0.860	0.823	0.818	0.682
		PS-70%	589	930.44	0	299.72	83.52	0.831	0.875	0.835	0.830	0.708
		PS-80%	392.67	471.11	11.78	205.73	79.88	0.816	0.848	0.799	0.796	0.649
	LK	UTS	1965.44	2920.89	47.11	1002.56	82.31	0.823	0.861	0.823	0.818	0.683
		CVF-10	1974.84	2873.78	94.22	993.16	81.68	0.817	0.849	0.817	0.813	0.665
		PS-70%	621	860	71	268	81.39	0.811	0.829	0.814	0.811	0.641
		PS-80%	392.67	471.11	11.78	205.73	79.88	0.816	0.848	0.799	0.796	0.649
	RBF	UTS	1659.97	2968	0	1308.03	77.96	0.780	0.847	0.780	0.768	0.623
		CVF-10	1847	2968	0	1120.57	81.12	0.811	0.863	0.811	0.804	0.672
		PS-70%	576.47	930.44	0	312.26	82.83	0.824	0.871	0.828	0.822	0.697
		PS-80%	328.97	482.89	0	269.44	75.08	0.775	0.840	0.751	0.742	0.594
	PUK	UTS	1963.35	2968	0	1004.65	83.07	0.831	0.874	0.831	0.826	0.703
		CVF-10	1961.78	2920.89	47.11	1006.22	82.25	0.823	0.860	0.823	0.818	0.682
		PS-70%	619.29	918.67	11.78	269.44	84.54	0.842	0.875	0.845	0.842	0.719
		PS-80%	401.02	447.56	35.33	197.38	78.47	0.798	0.819	0.785	0.783	0.605
LR		UTS	2050.03	2744.22	223.78	917.97	80.76	0.831	0.825	0.808	0.805	0.633
		CVF-10	2051.07	2708.89	259.11	916.93	80.18	0.815	0.817	0.802	0.799	0.619
		PS-70%	615.11	859.78	70.67	273.62	81.07	0.843	0.826	0.811	0.808	0.636
		PS-80%	402.07	424	58.89	196.33	76.39	0.804	0.788	0.764	0.763	0.553
C4.5		UTS	2043.24	2956.22	11.78	924.76	84.22	0.866	0.878	0.842	0.838	0.719
		CVF-10	2064.13	2779.56	188.44	903.87	81.59	0.833	0.835	0.816	0.813	0.651
		PS-70%	640.18	871.56	58.89	248.55	83.1	0.847	0.845	0.831	0.829	0.675
		PS-80%	414.08	412.22	70.67	184.33	76.41	0.825	0.781	0.764	0.764	0.545
NB		UTS	2079.27	2638.22	329.78	888.73	79.47	0.833	0.806	0.795	0.793	0.600
		CVF-10	2055.25	2650	318	912.75	79.26	0.823	0.805	0.793	0.791	0.597
		PS-70%	582.22	930.44	0	306.51	83.15	0.834	0.873	0.832	0.826	0.702
		PS-80%	405.2	412.22	70.67	193.2	75.59	0.810	0.775	0.756	0.756	0.532

Based on the results presented in Table 6, the application of ClassBalancer effectively enhances the overall performance of the models across various algorithms and kernel types. In particular, the SVM model with the PUK kernel (PS-70%) achieved the highest performance, with an accuracy of 84.54%, an AUC of 0.842, and an MCC of 0.719, indicating a well-balanced trade-off between accuracy and minority class detection capability. The C4.5 model also exhibited consistent and robust performance with relatively high accuracy across most test methods, while LR and NB achieved competitive but slightly lower results. These

findings suggest that, tree-based and non-linear kernel models tend to outperform linear models in recognizing complex attribute relationships.

The confusion matrix analysis indicates that the "not accepted" category (negative class) remains more challenging to classify or predict accurately than the "accepted" category. Higher FN rates in models like LR and C4.5 show difficulty in detecting minority instances, despite high TP and overall accuracy favoring the majority class. In contrast, PUK and NPK achieved balanced predictions with strong sensitivity on PS-70%, while the RBF kernel showed lower stability on PS-80% despite relatively high accuracy.

4. CONCLUSION

The SVM classification model was applied to a dataset comprising 5,936 records and four attributes in this study's testing process, which included three test configurations: use training set, CVF-10, PS of 70% and 80%. Before applying class balancing, the SVM-kernel model exhibited high accuracy but failed to generate any TP outcomes, indicating its inability to detect the minority class "not accepted" do to substantial class imbalance. This imbalance caused the model to become biased toward the majority class, resulting in a misleading perception of strong performance. Following the implementation of class balancing, the model's performance improved considerably, particularly in terms of AUC, F-measure, and MCC, indicating a more equitable classification between majority and minority classes. The SVM with PUK kernel and C4.5 models demonstrated the best results, suggesting that class balancing techniques effectively enhance model sensitivity and promote fairness in classification performance. This study is limited by the small number of attributes used, which may restrict the model's ability to capture complex patterns. Additionally, the absence of external validation and explicit analysis of overfitting risks may limit the generalizability of the results.

FUNDING INFORMATION

This research is self-funded, with all expenses covered solely by the research team.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Neni Purwati	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Windya Harieska Pramujati		✓				✓		✓	✓	✓	✓	✓		
A. Aviv Mahmudi	✓		✓	✓			✓			✓	✓		✓	✓
Mira Febriana Sesunan		✓				✓		✓	✓	✓	✓	✓		✓
Yahya	✓		✓	✓			✓			✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

INFORMED CONSENT

Not applicable, as this study did not involve human participants or personally identifiable information.

ETHICAL APPROVAL

Not applicable, as this study did not involve human or animal subjects and used only non-sensitive computational or public data.

DATA AVAILABILITY

The dataset underlying this study's findings is available from the corresponding author, [NP], upon request. Access is restricted and not publicly available to ensure the privacy and confidentiality of the research participants.

REFERENCES

[1] I. El Guabassi, Z. Bousalem, R. Marah, and A. Qazdar, "A recommender system for predicting students' admission to a graduate program using machine learning algorithms," *International Journal of Online and Biomedical Engineering*, vol. 17, no. 2, pp. 135–147, 2021, doi: 10.3991/ijoe.v17i02.20049.

[2] M. Valavala, W. Alhamdani, and H. Indukuri, "Expediting international student admission process using data analytics," *IOP Conference Series: Materials Science and Engineering*, vol. 1074, no. 1, 2021, doi: 10.1088/1757-899x/1074/1/012024.

[3] S. Ebiesuwa, T. Omolara, A. A. Yinka, O. O. Blaise, and A. Adio, "The need for domain expert in data mining application," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 3, pp. 639–650, 2021.

[4] M. Hinojosa, M. Alfaro, G. Fuertes, R. Ternero, P. Santander, and M. Vargas, "Optimizing university admission processes for improved educational administration through feature selection algorithms: a case study in engineering education," *Education Sciences*, vol. 15, no. 3, pp. 0–23, 2025, doi: 10.3390/educsci15030326.

[5] H. Saoud, A. Ghadi, and M. Ghailani, "Breast cancer diagnosis using machine learning and ensemble methods on large seer database," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 3, pp. 594–604, 2021.

[6] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Fair and transparent student admission prediction using machine learning models," *Algorithms*, vol. 17, no. 12, pp. 1–19, 2024, doi: 10.3390/a17120572.

[7] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Support vector machines in polymer science: a review," *Polymers*, vol. 17, no. 4, pp. 1–26, 2025, doi: 10.3390/polym17040491.

[8] H. ALMarwi and G. H. AL-Gaphari, "A review of feature selection methods in big data," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 4347–4366, Sep. 2024.

[9] Y. Altork, "Comparative analysis of machine learning models for wind speed forecasting: support vector machines, fine tree, and linear regression approaches," *International Journal of Thermofluids*, vol. 27, pp. 1–10, 2025, doi: 10.1016/j.ijft.2025.101217.

[10] A. Bisri, Supardi, Y. Heryatun, Hunainah, and A. Navira, "Educational data mining model using support vector machine for student academic performance evaluation," *Journal of Education and Learning*, vol. 19, no. 1, pp. 478–486, 2025, doi: 10.11591/edulearn.v19i1.21609.

[11] A. I. Gufroni, P. Purwanto, and F. Farikhin, "Academic performance prediction using supervised learning algorithms in university admission," *International Journal on Informatics Visualization*, vol. 9, no. 1, pp. 184–194, 2025, doi: 10.62527/joiv.9.1.2974.

[12] R. Wella, A. A. Sandra, and A. E. John, "Machine learning system for predicting student suitability for university courses," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 2, pp. 886–895, 2025, doi: 10.47709/brilliance.v5i2.5774.

[13] M. A. Mirza, S. B. Baba, Y. M. Aravind, S. Rahul, S. Nagaraju, and F. A. Mirza, "A novel approach for university admission prediction using feature selection and data upgrading strategies," *Journal of Advancement in Software Engineering and Testing*, vol. 8, no. 3, pp. 1–13, 2025, doi: 10.5281/zenodo.15405309.

[14] A. Kumar and D. Mishra, "Improving support vector machine using modified kernel function," *International Journal of Scientific Research and Modern Technology*, vol. 4, no. 5, pp. 1–5, 2025, doi: 10.38124/ijsrmt.v4i5.501.

[15] S. Pal, L. H. Trang, V. T. Hieu, D. D. Nguyen, D. Q. Vu, and I. Prakash, "Investigation of support vector machines with different kernel functions for prediction of compressive strength of concrete," *Journal of Science and Transport Technology*, vol. 4, no. 2, pp. 55–68, 2024, doi: 10.58845/jstt.utt.2024.en.4.2.55-68.

[16] D. Yu, L. Zhu, H. Shen, G. Tang, W. Hu, and F. Dong, "Classification method of surface defects of silicon nitride bearing rollers based on the coupling of convolutional neural network and dual-kernel support vector machine," *Materials Today Communications*, vol. 42, 2025, doi: 10.1016/j.mtcomm.2024.111337.

[17] K. Zub, P. Zhezhnych, and C. Strauss, "Two-stage PNN–SVM ensemble for higher education admission prediction," *Big Data and Cognitive Computing*, vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020083.

[18] N. A.-Tejera, M. Gamarra, J. I. Vélez, and E. Zurek, "A distance-based kernel for classification via support vector machines," *Frontiers in Artificial Intelligence*, vol. 7, pp. 1–15, 2024, doi: 10.3389/frai.2024.1287875.

[19] M. A. R. Canul, J. A. R.-Hernandez, A. Y. Alanis, J. C. G. Gomez, and J. Gálvez, "Modified soft margin optimal hyperplane algorithm for support vector machines applied to fault patterns and disease diagnosis," *Symmetry*, vol. 17, no. 10, pp. 1–40, 2025, doi: 10.3390/sym17101749.

[20] M. Koranga *et al.*, "SVM model to predict the water quality based on physicochemical parameters," *International Journal of Mathematical, Engineering and Management Sciences*, vol. 6, no. 2, pp. 645–659, 2021, doi: 10.33889/IJMEMS.2021.6.2.040.

[21] M. T. Sathe and A. C. Adamuthe, "Comparative study of supervised algorithms for prediction of students' performance," *International Journal of Modern Education and Computer Science*, vol. 13, no. 1, pp. 1–21, 2021, doi: 10.5815/ijmecs.2021.01.01.

[22] X. Cui, H. Liu, M. Fan, B. Ai, D. Ma, and F. Yang, "Seafloor habitat mapping using multibeam bathymetric and backscatter intensity multi-features SVM classification framework," *Applied Acoustics*, vol. 174, Mar. 2021, doi: 10.1016/j.apacoust.2020.107728.

[23] M. Ahmad, R. A. Al-Mansob, I. Jamil, M. A. Al-Zubi, M. M. S. Sabri, and A. C. Alguno, "Prediction of rockfill materials' shear strength using various kernel function-based regression models—a comparative perspective," *Materials*, vol. 15, no. 5, 2022, doi: 10.3390/ma15051739.

[24] WEKA, "Package weka.classifiers.functions.supportVector," *The University of Waikato, Hamilton,* 2022. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/supportVector/package-summary.html

[25] M. Hoss, "Checklist to transparently define test oracles for TP, FP, and FN objects in automated driving," 2023, *arXiv:2308.07106*

[26] K. T. Mori and S. Peters, "SHARD: safety and human performance analysis for requirements in detection," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 3010–3021, 2024, doi: 10.1109/TIV.2023.3320395.

[27] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, 2024, doi: 10.1038/s41598-024-56706-x.

[28] M. M. Bassiouni, R. K. Chakrabortty, K. M. Sallam, and O. K. Hussain, "Deep learning approaches to identify order status in a complex supply chain," *Expert Systems with Applications*, vol. 250, 2024, doi: 10.1016/j.eswa.2024.123947.
[29] R. A. Praba and L. Suganthi, "HARNet: automatic recognition of human activity from mobile health data using cnn and transfer learning of lstm with svm," *Automatika*, vol. 65, no. 1, pp. 167–178, 2024, doi: 10.1080/00051144.2023.2290736.

APPPENDIX

Table 1. Related work

Reference	Dataset	Method	Accuracy	Limitations	Analysis
Mirza *et al*. [13]	Combined data from various exams: GRE, TOEFL, EAMCET, NEET, and PGCET exam scores, Undergraduate GPA, University Rating, Gender, and Caste. Consisting of hundreds to thousands of records.	LR, DT, RF, k-nearest neighbors (KNN), SVM, and least absolute shrinkage and selection operator (LASSO); features selected using best subset selection and DT feature importance.	RF achieved the best performance (lowest mean absolute error (MAE) and root mean squared error (RMSE); no specific values provided).	Lack of external validation; dataset size unspecified; inclusion of sensitive attributes (gender, caste) may cause bias.	Emphasizes the importance of feature selection and ensemble methods but overlooks SVM kernel performance and the fairness aspect; future work should explore diverse SVM kernels to enhance accuracy and model fairness.
Kumar and Mishra [14]	The study uses multiple classification datasets from different domains, but lacks detailed information on dataset characteristics, limiting reproducibility.	The study employs a comparative experimental method to evaluate SVM with the proposed PRBF kernel against conventional kernels.	The polynomial RBF kernel outperforms RBF and Polynomial kernels in terms of classification accuracy.	The study is limited by the lack of detailed information on the dataset and parameters, as well as the absence of external validation and overfitting analysis.	The results suggest that the hybrid polynomial RBF kernel enhances SVM classification performance by capturing both local and global data patterns; however, limited methodological detail may restrict the generalizability of the findings.
Pal *et al*. [15]	The dataset consists of 236 samples obtained from the Red River surface water treatment plant project in Hanoi, Vietnam, with a split of 70%-30%.	SVM with linear (LIN), polynomial (POL), RBF, and sigmoid (SIG) kernels.	The SVM-RBF model demonstrates the best performance with an R value of 0.847, followed by SVM-POL, SVM-LIN, and SVM-SIG on the testing data.	The relatively small dataset size and the use of data sourced from a single project location limit the generalizability of the results to other conditions.	Kernel selection has a significant impact on SVM performance, with the RBF kernel demonstrating the highest effectiveness in modeling nonlinear relationships and being the most reliable among the tested kernels.
Yu *et al*. [16]	The dataset consists of images of silicon nitride bearing roller surfaces containing various types of surface defects.	A convolutional neural network (CNN) is combined with a dual-kernel support vector machine (DK-SVM), which leverages a combination of two kernels to enhance class separation capability.	The CNN–dual kernel SVM model demonstrates superior classification performance compared to conventional SVM or a standalone CNN.	The reliance on an industry-specific image dataset and the increased computational complexity resulting from integrating a CNN with a dual-kernel SVM.	The integration of CNN and dual-kernel SVM is effective in addressing image-based defect classification problems, as the CNN extracts representative features while the dual-kernel SVM enhances nonlinear class separation.
Zub *et al.* [17]	University admission data containing academic attributes (exam scores, GPA), demographic information (age, gender, nationality), and interview results; the total number of records is not specified.	DT, RF, SVM, LR, and neural network. Models were compared based on accuracy and interpretability.	RF and neural network achieved the best performance, while SVM produced competitive results depending on kernel parameter settings.	Dataset obtained from a single institution; sensitive attributes such as gender or race are not discussed; focus mainly on accuracy without fairness or efficiency evaluation.	The study introduces a probabilistic neural network (PNN)–SVM ensemble for admission prediction but is limited by a small dataset size and a lack of kernel selection analysis, restricting the model's generalizability and optimization.

BIOGRAPHIES OF AUTHORS

Neni Purwati is a lecturer in Medical Informatics, Faculty of Health Science, Universitas Muhammadiyah Lamongan, East Java, Indonesia. She holds a master's degree in Informatics Engineering with a specialization in Information Systems at the Institut Informatika dan Bisnis Darmajaya, from 2011 to 2013. She has been a lecturer since 2006 and has been nationally certified as a lecturer since 2016. Her research field is data analysis, including data warehouse, data mining, and data visualization. She once received a grant from the government as a beginner lecturer-researcher. She has also published books on data warehouse, data mining, fundamentals of mathematics in data science, and has several simple patents from several research studies she has conducted. She has authored or coauthored than 47 publications: 7 proceedings and 38 journals, with 9 9-H-index and 228 citations. As editor of several accredited journals, Sinta. She can be contacted at email: nenipurwati@umla.ac.id.

Windya Harieska Pramujati is a lecturer in the Department of Information Technology, Politeknik Negeri Malang, East Java, Indonesia. She has been a lecturer since 2023. She holds a bachelor's degree and a master's degree in the Department of Mathematics at Institut Teknologi Sepuluh Nopember, from 2016 to 2022, with a specialization in Applied Mathematics. She has experience as a research assistant in the Industrial and Financial Mathematics Laboratory, where she was involved in a research project on natural disaster insurance. Additionally, she has completed an internship at the Ministry of Finance of the Republic of Indonesia, conducting research on regression analysis. Her research fields are time series analysis and financial mathematics. She can be contacted at email: windya.harieska@polinema.ac.id.

A. Aviv Mahmudi completed his bachelor's degree in Information Systems at STMIK AKI Pati, graduating in 2009, and continued his studies in the Master of Information Systems program at Universitas Diponegoro Semarang, graduating in 2014. He is a lecturer at Universitas YPPI Rembang (formerly known as STIE YPPI Rembang). He actively writes articles for various national and international scientific journals/proceedings and is also active as an editor, reviewer, and peer reviewer for various research journals and community service journals. He has also contributed to several intellectual property rights and books. He is also active in the Alumni Association of the Indonesian Islamic Students Movement (IKA-PMII), a member of APTIKOM, the representative branch board of NU in Rembang District, the board of karang taruna in Rembang Regency, and the board of rembang supporters association (Ganster). He can be contacted at email: viva.althaf@gmail.com.

Mira Febriana Sesunan was still a bachelor (S1) student at STMIK Darmajaya, now at the Institut Informatika dan Bisnis Darmajaya, Department of Informatics Engineering while working as an assistant until she became a lecturer starting in 2003 at the campus, then in 2008 he continued his postgraduate (S2) program at Computer Science Study Program, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, and while working as a lecturer at STMIK Jenderal Achmad Yani; now Universitas Jenderal Achmad Yani Yogyakarta. She is currently a permanent lecturer in the Department of Information Systems, Faculty of Engineering, Universitas Darma Persada, East Jakarta. Several articles in SINTA nationally or internationally accredited journals and proceedings have been published, and will continue to publish other creative ideas. She can be contacted at email: mira_febriana@ft.unsada.ac.id.

Yahya interest in Computer Science began in 1997. This made him choose to enter Universitas Gunadarma, majoring in Informatics Engineering, and he successfully graduated in 2001. After graduating, he developed the knowledge he had gained into the world of work in the professional and educational fields. He then continued his education at the master's level in 2016 and completed his studies in 2018 with a master's in Computer Science Technology at Universitas Budi Luhur. He has an interest in application, web, and database development. He started his career as a lecturer in 2018 and became a permanent lecturer at the Department of Information Systems, Faculty of Engineering, Universitas Darma Persada from September 2019 until now. In addition, the author is also actively involved in community service in assisting MSMEs and writing books to make a positive contribution to the nation and state. He can be contacted at email: yahya@ft.unsada.ac.id.