Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 6 (2024) 1092

Journal of the Nigerian Society of Physical Sciences

On logistic regression versus support vectors machine using vaccination dataset

Olumide S. Adesina[a], Adedayo F. Adedotun[b,*], Kayode S. Adekeye[a], Ogbu F. Imaga[b], Adeleke J. Adeyiga[c], Toluwalase J. Akingbade[b]

[a]Department of Mathematical Sciences, Redeemer's University, Ede 232101, Osun State, Nigeria
[b]Department of Mathematics, Covenant University, Ota 11001, Ogun State, Nigeria
[c]Department of Computer Science and Information Technology, Bells University of Technology, Ota 11001, Ogun State, Nigeria

Abstract

The performance of two classification techniques, logistic regression and Support Vector Machines (SVMs), in assessing vaccination data is investigated in this study. The model was trained based on leave-out-one cross validation to obtain an accurate result. Simulated with ten thousand replications, a life data set was used to establish a better model. The findings from the simulation revealed that the logistic regression model slightly outperformed the SVM while the life data shows that the tuned SVM outperformed both the logistic and the SVM. This demonstrates the practical utility of advanced approaches such as SVMs in difficult categorization scenarios such as vaccination prediction. The study emphasizes the superiority of the customized SVM model in this setting, as well as the potential of machine learning approaches to increase comprehension of complicated healthcare scenarios and guide data-driven decision-making for influencing vaccination plans and public health. The study recommends the use of logistic regression if the data point is high.

DOI:10.46481/jnsps.2024.1092

Keywords: Logistic regression, Leave out-one cross validation, Support vectors machine, Prediction, Vaccination

Article History :
Received: 26 September 2022
Received in revised form: 28 November 2023
Accepted for publication: 23 December 2023
Published: 18 February 2024

Communicated by: T. Latunde

1. Introduction

Machine learning approaches have acquired substantial importance in a variety of fields in recent years, altering how judgments are made and patterns are discovered from complex datasets [1]. One example is the examination of medical and healthcare data, where machine learning algorithms play a critical role in revealing insights that aid in treatment optimization, disease prediction, and planning in public health. Among these methods, logistic regression and support vector machines (SVMs) are two popular strategies for classification tasks [2]. This study does a comparison of these two methods, specifically in the setting of a vaccine dataset. By exploring the advantages, limitations, and performance characteristics of both logistic regression and SVMs, this study aims to provide a comprehensive understanding of their applicability in the domain of vaccination prediction. The results of this research could con-

[☆] Only the first word and nouns should begin with a capital letter.
[*] Corresponding author: Tel.: +234-805-571-1272;
Email address: adedayo.adedotun@covenantuniversity.edu.ng (Adedayo F. Adedotun)

tribute to the advancement of data-driven decision-making in public health, potentially enhancing vaccination strategies and outcomes.

Machine learning has revolutionized the healthcare landscape by enabling data-driven insights for clinical decision making, disease prediction, and preventive intervention. Its capacity to analyze vast amounts of medical data with speed and accuracy has not only enhanced the precision of diagnoses and treatment plans but has also opened new frontiers in personalized medicine, paving the way for more targeted and effective healthcare strategies [3, 4]. With the increasing availability of electronic health records and vast medical datasets, machine learning algorithms have been employed to uncover hidden patterns, identify risk factors, and develop predictive models. Notably, classification algorithms such as logistic regression and support vector machines have demonstrated efficacy in analyzing medical datasets for predictive tasks. Machine learning has proven to enhance clinical decision-making processes by harnessing the wealth of patient data available. High accuracy in predicting heart failure hospitalizations based on electronic health records was achieved through the utilization of machine learning techniques [5]. The study conducted a rigorous and in-depth examination to assess the effectiveness and practicality of leveraging advanced deep-learning techniques for predicting patient mortality. This study aimed to uncover the capabilities and limitations of these cutting-edge methodologies within the healthcare domain, where accurate predictions of patient outcomes play an integral role in clinical decision-making, resource allocation, and ultimately, the delivery of high-quality healthcare services [6]. This capability serves as a valuable aid for healthcare professionals, aiding them in the early identification of high-risk patients who may benefit from specialized medical attention, as detailed in reference. Deep learning techniques were employed to predict sepsis onset [7], effectively leveraging the longitudinal nature of patient data to improve prediction accuracy. Machine learning techniques applied to historical healthcare data have revealed significant racial disparities in healthcare outcomes [8]. This crucial finding underscores the importance of addressing and mitigating these disparities to ensure equitable access and quality of healthcare services for all patient populations. The significance of model interpretability in the context of healthcare was underscored through a comprehensive comparison of different machine learning algorithms, with a specific focus on logistic regression [9]. This research emphasizes the critical need for healthcare practitioners to not only obtain accurate predictions but also to comprehend and trust the models' explanations, enabling informed and actionable decision-making in medical settings. The likelihood of a patient's readmission was calculated using support vector machines as a predictive tool. The support vector machines were found to be extremely beneficial as it helps to optimize resource allocation plans and results in patient care [10].

Machine learning has demonstrated its potential in predicting factors that influence vaccine uptake. Recent studies used machine learning algorithms to forecast the adoption of childhood vaccines by analyzing a variety of factors such as demographic information, socioeconomic status, and healthcare utilization data. This analytical approach not only enabled precise predictions but also the identification of vulnerable population groups [11]. Leveraging the capabilities of machine learning, the study forecasted the uptake of influenza vaccines among pregnant women, thereby assisting healthcare providers in customizing interventions to enhance vaccination rates. Additionally, the research explored the application of machine learning techniques in predicting vaccine hesitancy, a phenomenon that involves the postponement or refusal of vaccines, even when they are readily accessible and available for use [12]. The ability of machine learning techniques to analyze large datasets sourced from Twitter to gauge sentiment related to vaccine hesitancy was carried out. This analytical approach enabled the identification of prevalent public sentiments, but it also allowed for a more in-depth understanding of the underlying concerns and apprehensions associated with vaccine reluctance. The findings of this study have helped in developing public health communication strategies and interventions aimed at addressing vaccine hesitancy in communities [13]. Furthermore, machine learning methods were employed to predict the geographic spread of vaccine-related misinformation on social media platforms. The result of the research contributed to public health strategies and interventions [14]. Machine learning has also been employed in forecasting vaccination coverage based on socio-demographic factors. Machine learning models were utilized to forecast measles vaccination coverage in Pakistan, contributing to the optimization of resource allocation for vaccination campaigns [15]. To enhance the efficiency and precision of vaccination planning, the adoption of machine learning methodologies has been explored as a potential solution for forecasting vaccination coverage in settings characterized by resource constraints. This approach represents a significant step toward optimizing vaccination programs and ensuring that limited resources are deployed more strategically and effectively in the realm of public health [16].

While both logistic regression and Support Vector Machines (SVMs) have shown promise in various medical applications, their comparative performance in the specific context of vaccination prediction remains relatively unexplored. This research seeks to address this gap by conducting a head-to-head comparison of logistic regression and SVMs using a vaccination dataset. By evaluating factors such as prediction accuracy, interpretability, generalization ability, and robustness to different vaccination scenarios, this study aims to provide insights into which algorithm is better suited for predicting vaccination outcomes. The findings of this research could aid public health policymakers and practitioners in making informed decisions regarding vaccination strategies.

2. Materials and Methods

This section highlights the material and methods adopted in this study. A vaccination data set was used for the analysis, where the response variable is vaccination status (vaccinated, 1), (not vaccinated, 0). The predictor variables (VIR) stand for Vaccination based on immigration requirements, while VNA stands for Vaccination status based on vaccines not readily available

2.1. Logistic regression

Logistic regression model is of the form:

$$f(x_i) = In\left(\frac{\varphi_i}{1-\varphi_i}\right) = x_i\beta, \tag{1}$$

where

$$\varphi_i = \frac{e^{x_i\beta}}{1-e^{x_i\beta}}.$$

From Eq. (1), we have

$$f(x_i) = In\left(\frac{\varphi_i}{1-\varphi_i}\right) = x_i\beta.$$

The maximum likelihood function is given as

$$L = \sum_{i=1}^{n} y_i \log\left(\varphi_i\right) + \sum_{i=1}^{n} \left(1-y_i\right)\log\left(1-\varphi_i\right). \tag{2}$$

Taking partial derivative in Eq. (2), we have

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \varphi_i\right) x_i = 0. \tag{3}$$

Solution to Eq. (3) follows the intuition of iteratively reweighted least squares (IRLS). Therefore the maximum likelihood estimate of logistic model is

$$\hat{\beta}_{MLE} = T^{-1}X\prime\hat{G}\hat{z}, \tag{4}$$

where

$T = X'^{\hat{G}}X,\ \hat{G} = diag\left(\widehat{\varphi_i}\left(1-\widehat{\varphi_i}\right)\right)$ and $\hat{z} = log\left(\widehat{\varphi_i}\right) + \frac{y_i-\widehat{\varphi_i}}{\widehat{\varphi_i}\left(1-\widehat{\varphi_i}\right)}$.

Details on the estimation of the parameters of logistic regression can be found in Ref. [17].

2.2. The Support vectors machine

Boser, Guyon, and Vapnik created the Support Vector Machine (SVM) in 1992. SVM is an important machine learning algorithm for categorizing patterns. A classifier called Support Vector Machine was developed for binary classification. SVMs are used to solve classification issues in areas like pattern recognition and speech recognition because they outperform conventional machine learning techniques [18]. Because of its strong generalization capabilities and track record of high accuracy in training datasets, SVM stands out among other classification algorithms. The separation of data into different formats, which makes linear separation challenging, is one of the most challenging aspects of classification. The most challenging parts of using the SVM are picking the best kernel function and altering the SVM learning parameters [18]. The support vector machine technique requires the separation of data in the hyperplane. We first describe the maximal margin hyperplane. The maximal margin hyperplane is the solution to the optimization problem

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n} M, \tag{5}$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{6}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots, \beta_p x_{ip}\right) \geq M, \quad i = 1, \ldots, n. \tag{7}$$

The constraints (6)-(7) are to ensure that each observation is on the correct side of the hyperplane and at least a distance M from the hyperplane. Where M is the margin of the hyperplane, so, we seek to maximize M.

The resulting maximal margin hyperplane is not satisfactory because it has only a tiny margin. However, the support vectors classifier (SVC) is more robust to individual observations, and it better classifies most of the training observations. The SVC is the solution to the optimization problem

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n} M, \tag{8}$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 = 1. \tag{9}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots, \beta_p x_{ip}\right) \geq M\left(1-\ \epsilon_i\right), \tag{10}$$

$$\epsilon_i \geq 0,\ \sum_{i=1}^{n} \epsilon_i \leq C, \tag{11}$$

where M the width of the margin of the optimization problem, and the aim is to make M large as possible, and C is treated as a tuning parameter that is generally chosen via cross-validation.

The support vector machine (SVM) is an extension of the support classifier to include non-linear class boundary.

The maximal margin hyperplane depends directly on the support vectors. For computing the predictions, only the support vectors are involved, not the whole training set.

Support vector regression (SVR) is derived using one-dimensional form in Eq. (12).

$$y = f\left(x\right) = \sum_{i=1}^{M} w_i x_i + b,\ y, b\epsilon\mathbb{R},\ x, w \in \mathbb{R}^M. \tag{12}$$

For multidimensional data, we augment x by one and include b in the w vector to simplify the mathematical notation, and obtain a muultivariate regression in Eq. (13).

$$y = f\left(x\right) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b \quad x,\ w\epsilon\mathbb{R}^{M+1}. \tag{13}$$

Figure 1 as adopted from Ref. [19] shows the separation of the training data on the hyperplane

In Figure 1, the training observations are situated along the hyperplane, with blue dots representing the training sets residing in the upper portion of the hyperplane and red dots ($W^T X + b < 0$) representing those in the lower section ($W^T X + b > 0$).

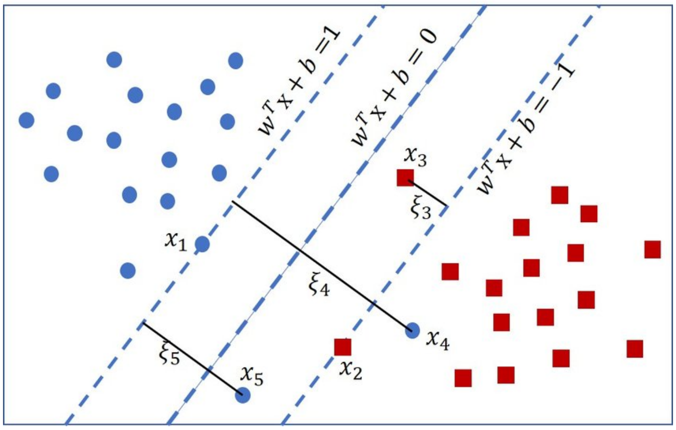

Figure 1. Support vectors classifier.

Table 1. Estimation error rate for simulated data.

Model	Logistic regression	SVM	Tuned SVM
MSE	0.2504391	0.3195852	0.2678139
RMSE	0.5004389	0.5653187	0.5175074

Table 2. Logistic regression output.

	Estimate	Std Error	Z-value	Pr(> |Z|)
Intercept	1.01750	0.41788	2.435	0.0149
VIR	-0.99238	0.19775	-5.018	5.21 $e^{-}07$
VNA	-0.06558	0.13299	-0.493	0.6219

Specifically, the support vectors, denoted by the points aligned with the broken lines, are the focal points for estimation. These support vectors are the ones that lie directly on the hyperplane, and they are selectively used for estimation, rather than involving the entire training dataset.

2.3. Simulation

Ten thousand (10000) binary response variable was simulated from binomial distribution Bin ≈ (1, 0.5), all x_s from uniform distribution as follows: $x_1 \approx (0, 1), x_2 \approx (0, 2), x_3 \approx (0, 1.5), x_4 \approx (0, 3), x_5 \approx (0, 1.8)$. The simulated data were fitted to both logistic model and the support vectors machine model using 80 % of the dataset as training set, and 20 % as testing set. The same was replicated for the life dataset.

3. Result

3.1. The simulation study

The results of the estimation based on simulated data are presented in Tables 1 - 4.

From the results of simulation in Table 1, logistic regression has lower cross-validation mean square error (MSE), and lower cross-validation root mean square (RMSE) than the Support Vectors machine (SVM) and the tuned support vector machine (Tuned SVM)

VIR in Table 2 stands for Vaccination based in immigration requirement, while VNA stands for Vaccination status is based on vaccines not readily available. The correlation coefficient of the variables is (-0.01187) which shows a very low negative correlation between the two variables, hence no indication of multicollinearity, it is further show using VIF plot in Figure 2.

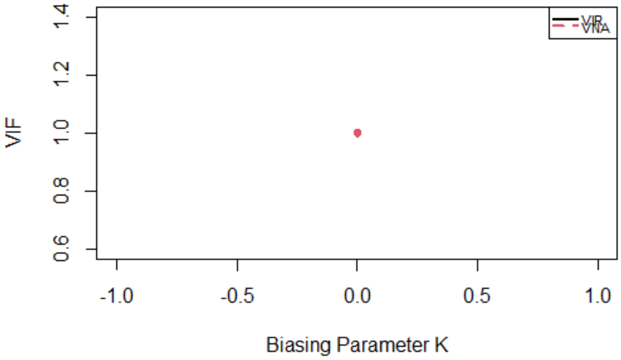

Figure 2. VIF plot for VIR and VNA.

Table 3. Estimation error rate for real-life data.

Model	Logistic regression	SVM	Tuned SVM
MSE	0.2099581	0.2210432	0.2031725
RMSE	0.4582119	0.4701523	0.4507466

Table 4. Support vector machine parameters.

Model	SVM	Tuned SVM
b	0.06187135	0.2524004
VIR	-5.154498	-4.091383
VNA	0.1456417	-0.2143416

In Figure 2, the red and black line overlapped at zero to form a dot. The presence of the red dot signifies that the biasing parameter for both variables is appropriately zero, and it also shows that the Variance Inflation Factor (VIF) of the two variables is around zero. This combined observation serves as a robust indicator, strongly suggesting the absence of multicollinearity within the variables under examination.

Table 3 shows that logistic regression has lower MSE, and RMSE than the support vectors machine while the tuned SVM estimation has a lower cross-validation mean square error (MSE), and lower cross-validation root mean square (RMSE) than the better than logistic and ordinary SVM, hence the tuned SVM is adjudged better.

From Table 4, the resulting equation from Eq. (13) is as follows:

$$\begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b = -4.091383 * VIR = d, \tag{14}$$

where $d = \pm 0.2143416 * VNA + 0.2524004$.

4. Conclusion

This research aimed to address the pertinent issue of choosing an appropriate classification model for analyzing vaccination data, ultimately comparing the performance of logistic regression and Support Vector Machines (SVMs). Through the application of machine learning methods, including cross-validation and utilizing both simulated and real-world vaccination datasets, this study shed light on the comparative efficacy of these algorithms. The results of this study notably favored the tuned SVM method, as it exhibited a lower error rate when predicting vaccination outcomes. This outcome underscores the significance of adopting advanced techniques like SVMs, particularly when dealing with complex classification tasks such as vaccination prediction. The findings emphasize the importance of considering algorithm-specific performance and fine-tuning for optimal results in healthcare-related predictive modeling.

The current study only compares the logistic regression and SVR, and the findings from the simulation revealed that the logistic regression model slightly outperformed the SVM while the life data shows that the tuned SVM outperformed both the logistic and the SVM, the better performance of SVM could be as a result of the large data set.

Future research could explore several directions to expand upon the current findings. Firstly, the inclusion of a wider spectrum of machine learning algorithms would provide a more comprehensive comparison and assist in identifying the strengths of various methods beyond logistic regression and SVMs. Based on the findings of this study, several recommendations emerge for both researchers and practitioners. For practitioners and policymakers, the findings underline the potential of SVMs, with proper parameter tuning, in enhancing vaccination decisions and administration strategies. Utilizing advanced machine learning methods to predict vaccination decisions could enable targeted interventions, optimize resource allocation, and ultimately contribute to improved public health outcomes.

In closing, this research not only contributes to the ongoing dialogue on machine learning in healthcare but also offers a pathway for future investigations in the realm of vaccination decision prediction and beyond. Through the intersection of data-driven methodologies and public health imperatives, the potential to reshape vaccination strategies for better preventive outcomes is both promising and inspiring.

Acknowledgment

The authors hereby acknowledge Covenant University Centre for Research, Innovation and Discovery (CUCRID) for their support toward the completion of this research.

References

[1] T. P. Iyiola, H. I. Okagbue, A. F. Adedotun & T. J. Akingbade, “The effects of decomposition of the goals scored in classifying the outcomes of five English Premier League seasons using machine learning models”, Advances and Applications in Statistics **87** (2023) 13. http://dx.doi.org/10.17654/0972361723026.

[2] O. S. Adesina, A. F. Adedotun, D. S. Oladepo & T. F. Adesina, “Knowledge, attitude, and perception of health and non-healthcare workers towards COVID-19 vaccination: Machine learning approach”, International Journal of Sustainable Development and Planning **17** (2022) 2015. https://doi.org/10.18280/ijsdp.170702.

[3] O. S. Adesina, A. F. Adedotun, N. O. Adeboye, T. F. Adesina, H. I. Okagbue & A. F. Gboyega, “On COVID-19 vaccination in Nigeria: An empirical study”, International Journal of Design & Nature and Ecodynamics **18** (2023) 219. https://doi.org/10.18280/ijdne.180128.

[4] M. H. Alsharif, A. H. Kelechi, K. Yahya & S. A. Chaudhry, “Machine learning algorithms for smart data analysis in internet of things environment: Taxonomies and research trends”, Symmetry **12** (2020) 88. https://doi.org/10.3390/sym12010088.

[5] C. T. Bauch, E. Szusz & L. P. Garrison, “Scheduling of measles vaccination in low-income countries: Projections of a dynamic model”, Vaccine **27** (2009) 4090. http://dx.doi.org/10.1016/j.vaccine.2009.04.079.

[6] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt & A. Esteva, “Scalable and accurate deep learning with electronic health records”, npj Digital Medicine **1** (2018) 1. http://dx.doi.org/10.1038/s41746-018-0029-1.

[7] J. Wiens, E. S. Shenoy & R. B. Parikh, “Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology”, Clinical Infectious Diseases **66** (2017) 149. http://dx.doi.org/10.1093/cid/cix731.

[8] Z. Obermeyer, B. Powers, C. Vogeli & S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations”, Science **366** (2019) 447. http://dx.doi.org/10.1126/science.aax2342.

[9] R. Caruana, Y. Lou, J. Gentry & G. Hooker, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”, In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining **2** (2015) 1721. http://dx.doi.org/10.1145/2783258.2788613.

[10] A. O. Eni, M. G. Soluade, P. E. Efekemo, P. Oghenevwairhe, T. T. Igwe & A. Olabode, “Poor knowledge of human Papilloma virus and vaccination among respondents from three Nigerian states”, Journal of Community Health **43** (2018) 1201. http://dx.doi.org/10.1007/s10900-018-0540-y.

[11] D. Rey, L. Fressard, S. Cortaredona, A. Bocquier & P. Peretti-Watel, “Vaccine hesitancy in the French population in 2016, and its association with vaccine uptake and perceived vaccine risk-benefit balance **23** (2016) 17. http://dx.doi.org/10.2807/1560-7917.ES.2018.23.17.17-00816.

[12] M. C. Nunes, S. Walaza, S. Meiring, H. J. Zar, G. Reubenson, M. McMorrow, S. Tempia, L. Rossi, R. Itzikowitz, K. Bishop, A. Mathunjwa, A. Wise, F. K. Treurnicht, O. Hellferscee, M. Laubscher, Serafin N, C. L. Cutland, S. A. Madhi & C. Cohen, “Effectiveness of Influenza vaccination of pregnant women for prevention of maternal and early infant Influenza-associated hospitalizations in South Africa: A prospective test-negative study”, Open Forum Infect Dis. **9** (2022) 11. https://doi.org/10.1093/ofid/ofac552.

[13] S. L. Wilson & C. Wiysonge, “Social-media and vaccine hesitancy working group”, BMJ Global Health **5** (2020) 1. http://dx.doi.org/10.1136/bmjgh-2020-004206.

[14] J. Pinchoff, J. Chipeta & G. C. Banda, “Spatial clustering of measles cases during endemic (1998–2002) and epidemic (2010) periods in Lusaka, Zambia”, BMC Infect Dis. **15** (2015) 121. https://doi.org/10.1186/s12879-015-0842-y.

[15] I. U. Rehman, A. Bukhsh & T. M. Khan, “Measles in Pakistan: Time to make steps towards eradication”, Travel Med Infect Dis. **18** (2017) 67. https://doi.org/10.1016/j.tmaid.2017.08.002.

[16] T. M. Lincoln, B. Schlier & F. Strakeljahn, “Taking a machine learning approach to optimize prediction of vaccine hesitancy in high income countries”, Sci Rep. **12** (2022) 2055. https://doi.org/10.1038/s41598-022-05915-3.

[17] F. A. Awwad, K. A. Odeniyi, I. Dawoud, Z. Y. Algamal, M. R. Abonazel, B. M. G. Kibria & E. T. Eldin, “New Two-Parameter estimators for the logistic regression model with multicollinearity”, WSEAS Transactions on Mathematics **21** (2022) 403. https://doi.org/10.37394/23206.2022.21.48.

[18] S. A. Abdulraheem, S. Aliyu & F. B. Abdullahi, “Hyper-parameter tuning for support vector machine using an improved cat swarm optimization algorithm”, Journal of the Nigerian Society of Physical Sciences **4** (2023) 1007. https://doi.org/10.46481/jnsps.2023.1007.

[19] H. A. Zisserman, “Machine Learning Lecture 2: The SVM classifier”, 2015. https://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf.