International Journal of Knowledge Based Computer Systems
10 (1) June 2022, 33-39
http://www.publishingindia.com/ijkbcs/

A Comprehensive Survey on Support Vector Machines for Intrusion Detection System

Akram Salim Khanfar[1], Firdous Ahmad Lone[2] and MD Moizuddin[3*]
[1]Faculty, King Saud University, Saudi Arabia. Email: akram.khanfar2011@gmail.com
[2]Faculty, King Saud University, Saudi Arabia. Email: lonefirdous686@gmail.com
[3]Faculty, King Saud University, Saudi Arabia. Email: moizqa12@gmail.com
*Corresponding Author

Abstract: Machine learning is a widely interdisciplinary field centered on theories from cognitive science, computer science, statistics, optimization and many other theoretical and mathematical disciplines. Classification is a supervised learning technique used in machine learning to evaluate a given dataset and to create a model that divides data into a desired and distinct number of groups. The strength of SVMs lies in their use of nonlinear kernel features that map input into high-dimensional spaces of features implicitly. We'll address the value of SVMs in this survey article. Discussing their SVM tuning parameters as well. The main purpose of this paper is to include detailed studies on SVM implementations by contrasting the current ML models with the SVM versions, also poses the problems of the intrusion detection method of the support vector machines, and also this paper provides researchers with a summary of the SVM that assists in their future analysis.

Keywords: Data Mining (DM), Intrusion Detection System (IDS), Machine Learning (ML), Optimization, Support Vector Machines (SVMs).

I. Introduction

The Machine learning is a branch of Artificial Intelligence, and the development of techniques, methods, and algorithms is an evolving field. These algorithms allow machines to understand the processes, assignments, and decisions that are made. In Machine learning, classification is supervised approach and using the classification techniques helps to classify the data into classes so that it can easily be identified. Classification is one of the techniques for machine learning that allows to group data to extract characteristics and forecast future effects. There are plenty of machine learning various algorithms in order to distinguish the data into classes. Support Vector Machine is among the most popular techniques for constructing models for machine learning. With less power to store, it has tremendous accuracy.

Machine learning (ML) algorithms for computational solutions can be used to perform efficiently. In recent years, there have been plenty of work carried out in the field of Support vector machines (SVM). SVM have shown good results in classification, generalization performance on many problems.

This paper is organized as follows: Section II presents the SVM's associated work as well as the SVM's findings and Section III presents Support Vector Machine and also covers the importance of SVM, SVM's tuning parameters. In terms of the Intrusion Detection Method, Section IV outlines the study of different SVM implementations. The SVM challenges are outlined in Section V.

Related Work

SVM is one of the powerful supervised methods for solving the classification and regression problems and also to provide the optimal solution [1] [2]. Through each point, it has the capacity to solve even the shortest classification problem. SVM offers solutions to problems relating to the fitting of training problems within a personal computer or workstation's storage capacity, since SVM does not need any matrix equations and is less likely to have issues with numerical formats [2]. SVM is an effective tool for target object selection and detection in a medical imaging device and microcalcifications [3]. SVM provides the best results on application to face detection and reuters collection and also give the new technique for implementing the SVM algorithm efficiently. SVMs have attractive qualities, such as classification accuracy, computational models, simple geometric interpretation, and stronger intrusion detection efficiency [4]. In classification methods, the performance of the SVM depends on understanding the necessary parameters and soft-margin coefficient of the kernel function [5]. The previous studies showed that, relative to linear SVM algorithms that use a single CPU, parallel SVMs can reach large speeds [6]. SVMs have been extended to several machine learning tasks. From an adequate set of kernel functions, it creates learning frameworks and architectures [7]. By introducing the information geometric of Riemannian geometry structure induced by the kernel, it is possible to improve the SVMs classifier performance [7]. With Dynamic Time Warping distance measurement, an

improved SVM scheme was used as a feature for the SVM Classifier [8] [9]. SVM classifiers are used in many fields for text classification [10], facial components detection and tracks facial and emotional expression recognition [11] [12]. SVM identify sets of genes with a common function using expression data analysis [13] and also provides best performance compared to Parzen windows and Fisher's linear discriminant [14]. SVMs were used to conduct malfunction categorization [15]. With a rule-based decision tree (RBDT), a multi-class support vector machine classification model is used to identify the faults of water quality sensors due to its reliability and generalization [16] [17].

II. Support Vector Machine

SVM is a supervised algorithm. The key goals of SVM are classification and regression. It's based on the concept of statistics and Vapnik-Chervonenkis dimensions [18] [19] [20]. The main purpose is to find the optimal hyperplane by dividing the data points into two components and maximizing the margin, in this way it solves the classification and regression problems. In 2D, hyperplane is line and in 3D it is a plane (also called n-dimensional line) [18]. This process contains data from training and testing data from research. The algorithm generates an optimal hyperplane in training data (supervised learning), which categorizes new instances and then the evaluation process is carried out from the constructed model. SVM takes minimal training area and less processing time and also avoids overfitting problems [21] [18].

Suppose we have two groups of labels and they are plotted on a graph as seen in Fig. 1 below.

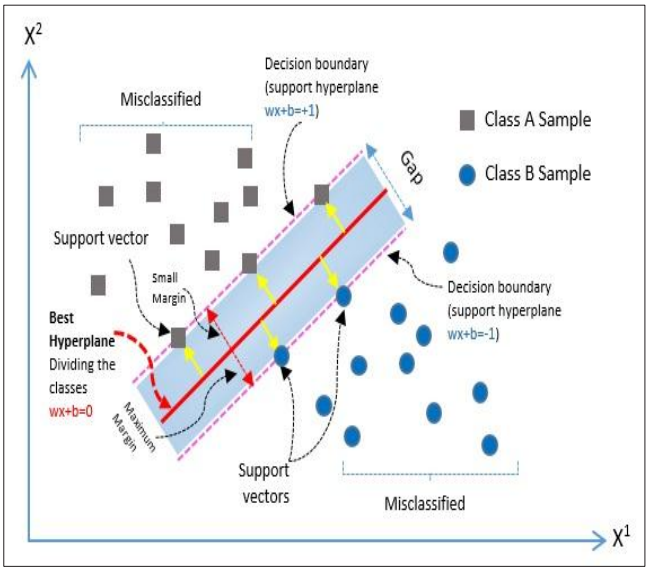

Fig. 1: The Concept of SVM-Hyperplane

Mathematically, for a hyperplane which divides the given data into two classes is represented by an equation as follows:

$$wx + b = 0 \quad (22)$$

A. Importance of Support Vector Machines

SVMs have been shown to work well, like analysis of remote protein sequence homology and recognition of gene transcription sites, it is used to examine expression data. It helps to overcome the challenges caused by high-dimensional data on gene expression data. One explanation for their effectiveness is that SVMs are targeted at minimizing error rates directly.

B. SVM Tuning Parameters

The categorization of groups in real-world situations is time-consuming. In order to overcome this, we use the parameters such as kernel, regularization, gamma, margin which are called tuning parameters or varying parameters.

C. Kernel

It is the backbone of SVM algorithm. It can resolve any complicated task. The SVM uses kernel-defined mathematical functions. These functions, such as linear, non-linear, polynomial, radial and sigmoid, take the data input training set and transform a non-linear decision into a high-dimensional space linear equation. In general, 2D data is converted into 3D data, as shown in Fig. 2.

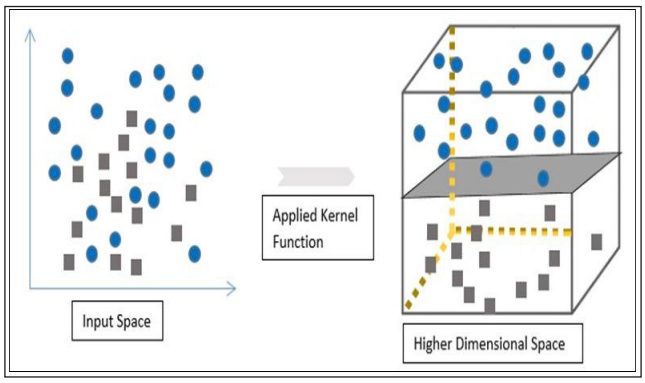

Fig. 2: Applied Kernel Function to Input Space

Mathematically, Kernel Function Equation is represented as:

$$K(x)= 1, \text{ if } ||x||<=1 \quad (22)$$

$$K(x)=0, \text{ Otherwise} \quad (22)$$

D. Regularization

Regularization is a computational approach that seeks to make a model structure as easy as possible to create. The optimized model will minimize the implications of overfitting at a small cost of performance. It is a way to reduce the model's complexity by penalizing the overfitting loss function. In mathematical

terms it is adding the sum of the weights to the cost function and also represented as:

$$J = \frac{1}{2}\sum_{i=1}^{M}(d_i - y_i)^2 + \lambda\frac{1}{2}\|w\|^2 \quad (17)$$

E. Gamma

Gamma is a kernel parameter of the RBF. It is known as the decision boundary. The greater Gamma value means more curvature, the decision region is high, and less curvature is indicated by the lower value, and the decision region is smaller. The gamma equation is mathematically expressed as:

$$\gamma = -\frac{1}{2\sigma^2} \quad (23)$$

F. Margin

It is the distance between the decision boundary to the closest data point in the given set of class. Models with higher margins yields better classification and performance.

G. Types of SVM Classifier

Support vectors are actually the coordinates that plots each data item as points, which are closest to the hyperplane. SVM choses the extreme points and these extreme points are called support vectors. There are two type of SVM Classifiers: Linear SVM, Non-linear SVM as shown in Fig. 3.

Linear SVM: In linear SVM, the data points are classified into two classes by a straight line, which divides the two classes. In this the data is linearly arranged and can be easily separated by a straight line. In linear, we use x and y as dimensions.

Non-Linear SVM: We cannot separate the non-linear data by a straight line, so to divide non-linear data a third dimension z is required.

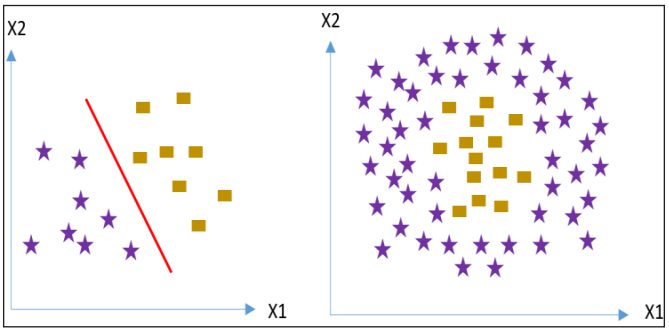

Fig. 3: Linear & Non-Linear SVMs (Ghosh et al., 2019)

III. Some of the Related Work of SVM in Terms of Intrusion Detection System

There are several SVM implementations using datasets. A list of the SVM implementations most used as seen in Table I. In the Table I, highlights the work carried out during the year 2016 to 2020.

Table I: Implementations of SVM in Terms of Intrusion Detection System

Challenges	Addressed by
Shortcomings related to the accuracy, number of selected features, and execution time	(Safaldin *et al*., 2020)
Managing changing data	(Jackson, 2002)
Low accuracy in weighted majority voting (WMO) approach	(Aburomman & Ibne Reaz, 2016)
Need complex features such as multi-classifier and feature selection in IDS	(Yang *et al*., 2016)
Lead to a long detection delay in the practical application scenario	(Gao *et al*., 2019)
Does not give the detailed information on the structure and characteristics of the malware	(Vinayakumar *et al*., 2019)
False positive rate	(da Costa *et al*., 2019)
Time cost in the data optimization stage and support for online processing	(Ren *et al*., 2019)
Large scale network will require additional infrastructure	(Taher *et al*., 2019)
High dimensionality of problems	(Tavara, 2019)
Address severe class imbalance, network traffic variability	(Gu & Lu, 2020)
Only a few numbers of works have been designed for detecting anomalies in the hosts	(Hosseinzadeh *et al*., 2020)

Challenges	Addressed by
Curse of dimensionality, Irrelevant features	(Hosseinzadeh *et al*., 2020)
Bolt loosening detection	(F. Wang *et al*., 2020)
Algorithmic complexity	(Cervantes *et al*., 2020)
Development of optimal classifiers for multi-class problems	(Cervantes *et al*., 2020)
The selection of kernel for a problem, Choosing good quality kernel parameters	(Nayak *et al*., 2015)
Speed and size in training and testing, training for very large datasets	(Tavara, 2019)
There are no theories concerning how to choose good kernel functions in a data-dependent way	(Byun & Lee, 2002)
Parallel algorithmic approaches for implementation of SVMs	(Tavara, 2019)
Exceedingly high time complexity in DTW computation	(Thapanan Janyalikit, 2016)

Conducted a search of the keyword 'Support Vector Machines' on numerous search engines like IEEE, Google Scholar, Elsevier, and Springer and the result shown in the below Table II.

Table II: Searched Keyword Support Vector Machines

Support Vector Machines	Results
Google Scholar	2,200,000
IEEE	63,369
Elsevier	8,896
Springer	24,737

IV. Challenges of SVM

Many researchers have noted a number of challenges in data mining science. Some of these are shown in Table III and need further focus from study.

Table III: Implementations of SVM in Terms of Intrusion Detection System

Authors & Year	Proposed Model/Method	Dataset Used	AD	FS	E. Criteria
(Yang *et al*., 2016)	ICPSO-SVM	KDD Cup 1999	Yes	Yes	FPR, DR
(Rebai, 2016)	ML-MKL	UCI dataset COIL-20	Yes	Yes	Accuracy, TP, TN
(Aburom man & Reaz, 2017)	LDA-PCA	KDD99	Yes	Yes	Accuracy, FP
(Liang *et al*., 2019)	Clustering-SVM Ensemble Method	NSL-KDD	Yes	Yes	Accuracy, Time, DR, FAR
(Safaldin *et al*., 2020)	GWOSVM-IDS	NSL-KDD	Yes	Yes	Accuracy, No. of features, Time, FR, DR
(H. Wang *et al*., 2017)	LMDRT-SVM	Gure-KDD dataset	Yes	Yes	Accuracy
(Al-Qatf *et al*., 2018)	STL-IDS	NSL-KDD	Yes	Yes	Accuracy
(Gu *et al*., 2019)	DT-EnSVM	NSL-KDD	Yes	Yes	Accuracy, DR, FAR
(Saleh *et al*., 2019)	Hybrid IDS	KDD Cup 1999, NSL-KDD, Kyoto 2006+ dataset	Yes	Yes	DR, Sensitivity, Specificity, Precision
(Tao et al., 2018)	FWP-SVM-GA	KDD Cup 1999	Yes	Yes	FPR, FNR, DR, Accuracy
(Kabir et al., 2018)	OA-LS-SVM	KDD Cup 1999	Yes	Yes	Accuracy, FAR
(Kavitha & Elango, 2020)	GRRF-FWSVM	KDD Cup 1999	Yes	Yes	Precision Recall F-Score
(Al Shorman et al., 2020) the number of Internet of Things (IoT	GWO-OCSVM	N-BaIoT	Yes	Yes	Average detection time, TPR, FPR

Authors & Year	Proposed Model/Method	Dataset Used	AD	FS	E. Criteria
(Roopa Devi & Suganthe, 2020)	HGWCSO with ETSVM	NSL-KDD	Yes	Yes	Precision, Recall, Sensitivity, Specificity, Accuracy
(Roopa Devi & Suganthe, 2020)	ROC and Confusion Matrix	NSL-KDD	Yes	Yes	Accuracy, Error, time
(Mighan & Kahani, 2020)	hybrid SAE–SVM scheme	NSK KDD, Kyoto, CDMC 2012, KDD Cup 1999 and UNB ISCX 2012	Yes	Yes	Accuracy, Recall, Time, Precision, F-measure
(Kumar & Ramasamy, 2020)	CSO-SVM algorithm	NSL-KDD	Yes	Yes	Accuracy, Recall Sensitivity Specificity Precision
(Ye *et al*., 2019)	GOA-SVM	KDD Cup, different datasets	Yes	Yes	Time, Accuracy

V. Conclusion and Future work

SVM are based on the concept of statistical learning theory. In SVM the inputs are placed in 3D space, where the different class groups are mapped using the mathematical functions. For each parameter SVM represents an optimal solution. This SVM algorithm is different from the other algorithms in terms working and in mapping the inputs on space. Kernel is important parameter on which this SVM works. In this paper, we have discussed the related works of SVM and highlighted the challenges of SVM, which will be helpful for the researchers in their future work. Lots of research work is ongoing to extend the scope and increase the performance and accuracy in detecting the Intrusions in a network. In future, we will focus on SVM applications and comparison of SVM methods with other machine learning techniques.

References

[1] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, 2020, doi: 10.1016/j.neucom.2019.10.118.

[2] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," *Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2019*, no. Iciss, 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.

[3] Q. Yang, H. Fu, and T. Zhu, "An optimization method for parameters of SVM in network intrusion detection system," *Proc. - 12th Annu. Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2016*, 2016, pp. 136-142, doi: 10.1109/DCOSS.2016.48.

[4] I. Rebai, Y. BenAyed, and W. Mahdi, "Deep kernel-SVM network," *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 2016, pp. 1955-1960.

[5] A. A. Aburomman, and M. B. I. Reaz, "Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection," *Proc. 2016 IEEE Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. IMCEC 2016*, 2016, pp. 636-640, doi: 10.1109/IMCEC.2016.7867287.

[6] D. Liang, Q. Liu, B. Zhao, Z. Zhu, and D. Liu, "A clustering-SVM ensemble method for intrusion detection system," *2019 8th Int. Symp. Next Gener. Electron. ISNE 2019*, vol. 2, no. 2, pp. 1-3, 2019, doi: 10.1109/ISNE.2019.8896514.

[7] M. Safaldin, M. Otair, and L. Abualigah, "Improved binary gray wolf optimizer and SVM for intrusion detection system in wireless sensor networks," *J. Ambient Intell. Humaniz. Comput.*, 2020, Art. no. 0123456789, doi: 10.1007/s12652-020-02228-z.

[8] H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," *Knowledge-Based Syst.*, vol. 136, pp. 130-139, 2017, doi: 10.1016/j.knosys.2017.09.014.

[9] M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, "Deep learning approach combining sparse autoencoder with SVM for network intrusion detection," *IEEE Access*, vol. 6, no. c, pp. 52843-52856, 2018, doi: 10.1109/ACCESS.2018.2869577.

[10] J. Gu, L. Wang, H. Wang, and S. Wang, "A novel approach to intrusion detection using SVM ensemble with feature augmentation," *Comput. Secur.*, vol. 86, pp. 53-62, 2019, doi: 10.1016/j.cose.2019.05.022.

[11] A. I. Saleh, F. M. Talaat, and L. M. Labib, "A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers," *Artif. Intell. Rev.*, vol. 51, no. 3, pp. 403-443, 2019, doi: 10.1007/s10462-017-9567-1.

[12] P. Tao, Z. Sun, and Z. Sun, "An improved intrusion detection algorithm based on GA and SVM," *IEEE Access*, vol. 6, pp. 13624-13631, 2018, doi: 10.1109/ACCESS.2018.2810198.

[13] E. Kabir, J. Hu, H. Wang, and G. Zhuo, "A novel statistical technique for intrusion detection systems," *Futur. Gener. Comput. Syst.*, vol. 79, pp. 303-318, 2018, doi: 10.1016/j.future.2017.01.029.

[14] G. Kavitha, and N. M. Elango, "An approach to feature selection in intrusion detection systems using machine learning algorithms," *Int. J. e-Collaboration*, vol. 16, no. 4, pp. 48-58, 2020, doi: 10.4018/IJeC.2020100104.

[15] A. Al Shorman, H. Faris, and I. Aljarah, "Unsupervised intelligent system based on one class support vector machine and grey wolf optimization for IoT botnet detection," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 7, pp. 2809-2825, 2020, doi: 10.1007/s12652-019-01387-y.

[16] E. M. Roopa Devi, and R. C. Suganthe, "Enhanced transductive support vector machine classification with grey wolf optimizer cuckoo search optimization for intrusion detection system," *Concurr. Comput.*, vol. 32, no. 4, pp. 1-11, 2020, doi: 10.1002/cpe.4999.

[17] S. N. Mighan, and M. Kahani, "A novel scalable intrusion detection system based on deep learning," *Int. J. Inf. Secur.*, 2020, Art. no. 0123456789, doi: 10.1007/s10207-020-00508-5.

[18] D. V. Kumar, and V. Ramasamy, "Improved intrusion detection classifier using cuckoo search optimization with support vector machine," *ICTACT J. Soft Comput.*, vol. 10, no. 2, pp. 2029-2034, 2020, doi: 10.21917/ijsc.2020.0287.

[19] Z. Ye, Y. Sun, S. Sun, S. Zhan, H. Yu, and Q. Yao, "Research on network intrusion detection based on support vector machine optimized with grasshopper optimization algorithm," *Proc. 2019 10th IEEE Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS 2019*, vol. 1, pp. 378-383, 2019, doi: 10.1109/IDAACS.2019.8924234.

[20] J. Jackson, "Data mining; A conceptual overview," *Commun. Assoc. Inf. Syst.*, vol. 8, 2002, doi: 10.17705/1cais.00819.

[21] A. A. Aburomman, and M. B. I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Appl. Soft Comput. J.*, vol. 38, pp. 360-372, 2016, doi: 10.1016/j.asoc.2015.10.011.

[22] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512-82521, 2019, doi: 10.1109/ACCESS.2019.2923640.

[23] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525-41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[24] K. A. P. da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, "Internet of things: A survey on machine learning-based intrusion detection approaches," *Comput. Networks*, vol. 151, pp. 147-157, 2019, doi: 10.1016/j.comnet.2019.01.023.

[25] J. Ren, J. Guo, W. Qian, H. Yuan, X. Hao, and H. Jingjing, "Building an effective intrusion detection system by using hybrid data optimization based on machine learning algorithms," *Secur. Commun. Networks*, vol. 2019, 2019, doi: 10.1155/2019/7130868.

[26] K. A. Taher, B. M. Y. Jisan, and M. M. Rahman, "Network intrusion detection using supervised machine learning technique with feature selection," *1st Int. Conf. Robot. Electr. Signal Process. Tech. ICREST 2019*, 2019, pp. 643-646, doi: 10.1109/ICREST.2019.8644161.

[27] S. Tavara, "Parallel computing of support vector machines: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, 2019, doi: 10.1145/3280989.

[28] J. Gu, and S. Lu, "An effective intrusion detection approach using SVM with Naive Bayes feature embedding," *Comput. Secur.*, vol. 103, 2021, Art. no. 102158, doi: 10.1016/j.cose.2020.102158.

[29] M. Hosseinzadeh, A. M. Rahmani, B. Vo, M. Bidaki, M. Masdari, and M. Zangakani, "Improving security using SVM-based anomaly detection: Issues and challenges," *Soft Comput.*, vol. 25, pp. 3195-3223, 2021, doi: 10.1007/s00500-020-05373-x.

[30] F. Wang, Z. Chen, and G. Song, "Monitoring of multi-bolt connection looseness using entropy-based active sensing and genetic algorithm-based least square support vector machine," *Mech. Syst. Signal Process.*, vol. 136, 2020, Art. no. 106507, doi: 10.1016/j.ymssp.2019.106507.

[31] J. Nayak, B. Naik, and H. S. Behera, "A comprehensive survey on support vector machine in data mining tasks: Applications & challenges," *Int. J. Database Theory Appl.*, vol. 8, no. 1, pp. 169-186, 2015, doi: 10.14257/ijdta.2015.8.1.18.

[32] H. Byun, and S. W. Lee, "Applications of support vector machines for pattern recognition: A survey," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2388, pp. 213-236, 2002, doi: 10.1007/3-540-45665-1_17.

[33] P. S. Thapanan Janyalikit, "Intelligent information and database systems: 8th Asian conference, ACIIDS 2016 da Nang, Vietnam, March 14-16, 2016 proceedings, Part I," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9621, pp. 616-625, 2016, doi: 10.1007/978-3-662-49381-6.