ICIC Express Letters
Volume **17**, Number **4**, April **2023**

ICIC International ©2023 ISSN 1881-803X
pp. **471–478**

VECTORIZATION METHOD BASED ON HIGH CORRELATION FOR MULTIVARIATE TIME SERIES HYBRID FILTER WRAPPER FEATURE SELECTION

Ani Dijah Rahajoe[1,*], Edi Winarko[2] and Suryo Guritno[2]

[1]Department of Computer Science
Universitas Pembangunan Nasional Veteran Jawa Timur
Raya Rungkut Madya, Gunung Anyar, Surabaya 60294, Indonesia
*Corresponding author: anidijah.if@upnjatim.ac.id

[2]Department of Computer Science and Electronics
Universitas Gadjah Mada
Bulaksumur Yogyakarta 55281, Indonesia
{ ewinarko; suryoguritno }@ugm.ac.id

Received March 2022; accepted June 2022

Abstract. *One of the techniques to reduce the multivariate time series dimension is to transform each Multivariate Time Series (MTS) dataset into a single row or column called vectorization. This paper contributes to using a new method in forming vectorization based on principal component analysis through an observation time analysis factor of each multivariate time series data without removing any information from the original data. The vectorization method is called Vectorization for Time of Observation Based on High Correlation (VecTOR), which is included in the filter method for feature selection. The wrapper method selects variables from the vectorization matrix with the Genetic Algorithm – Support Vector Machine algorithm (GASVM). VecTOR-GASVM is compared to four other methods: VecTOR – Support Vector Machine (VecTOR-SVM), VecTOR-GABayes, VecTOR Forward-Bayes, and VecTOR Backward-Bayes. The proposed method has been tested on the CMU and Wafer datasets. Results have shown that the feature selection of hybrid filter wrapper VecTOR has fewer features with the highest accuracy compared to the other four methods. In CMU data, the VecTOR-GASVM method has an accuracy of 100 per cent with 11 features selected. For the Wafer set of data, VecTOR-GASVM has an accuracy of 97.98 per cent with 2 features selected.*
Keywords: Vectorization, Support vector machine, Wrapper, Filter, Genetic algorithm

1. **Introduction.** Feature selection is used to reduce MTS data [1]. In [2], each row of 39 channel EEG data was encoded by an Autoregressive (AR) model of order 3, resulting in a 117-dimensional vector. The pre-processing process transforms the data information by looking at the correlation between variables using RFE (Recursive Feature Elimination) and testing using SVM [3]. This research confirms that the data should be a vector if using SVM. Support Vector Machine – Recursive Feature Elimination (SVM-RFE) on the MTS dataset, each MTS data matrix in the set must be first transformed into one row or column vector while retaining the correspondence to the original features. This process is called vectorization [3].

The method proposed is new in forming vectorization. This research is the perfection of the previous research on using the Feature Selection Based on Loadings Factor (FSBLF) method [4]. The formed vectorization is taken only from the MTS set of data's observation time. Thus, there is a matrix $P \times s$ where P is the number of observations (subject) and s is the number of selected variables. The data from s is the observation time of variable chosen from the most influential data from each MTS set, unlike other researchers whose

DOI: 10.24507/icicel.17.04.471

data are extracted. The method in this research is by selecting observation time twice in each data time series until a one-time observation for each MTS is obtained. The idea is to look for one most influential observation time from each MTS and create its vectorization column. This method is named Vectorization for Time of Observation Based on High Correlation (VecTOR). This high correlation is based on the factor analysis technique in classifying the data to form its principal component. Our previous research selected the observation time in each feature or FSBLF variable to have the observation time selection result [4]. Then, vectorization is done for each variable. Then, a ranking process is conducted to find out the variables whose accuracy level is the least. This research develops an algorithm by forming several vectorizations of MTS data into a vectorization matrix. Hence, only one observation time is needed for each MTS data for the following data mining process: prediction.

This research uses the wrapper method, which is GASVM because its accuracy level is better than the filter method [5]. GASVM is chosen because research using the GASVM wrapper method is better than other methods such as Particle Swarm Optimization – Support Vector Machine (PSO-SVM), though we know that its computation time is longer than the filter method [6]. To comprehend the performance of GASVM, more comparisons to other wrapper methods such as GA-Bayes, Forward-Bayes, and Backward-Bayes, are conducted. The results show that VecTOR-GASVM has the least selection features with 100 per cent accuracy on the CMU dataset. In the Wafer dataset, the VecTOR GASVM has 2 selected features with an accuracy of 97.98 per cent. The academic contribution of this research is a vectorized matrix for multivariate time series data. This vectorization matrix called VecTOR will make it easier for other researchers to select features for data in the form of multivariate time series. The remainder of this paper is organized as follows. In Section 2, the method is discussed. Section 3 discusses the results. Conclusions are presented in Section 4.

2. **Genetic Algorithm – Support Vector Machine (GASVM).** Genetic algorithm – support vector machine is one of the feature selection algorithms in the wrapper method category. GA was designed to optimally solve sequential decision processes more than to perform function optimization, but over the years, it has been used widely in both learning and optimization problems [7]. GASVM has differences in terms of the application of genetic algorithms for optimization. The genetic algorithm in GASVM has the task of randomly searching for chromosomes and representing features or variables. Therefore, the population group in one generation of the chromosome collects selected and non-selected variables. The selected chromosome is encoded with one, and the non-selected is encoded with zero. The chromosomes containing the selected features or variables are then used for the classification process using SVM to determine their accuracy. Figure 1 shows the overall GASVM algorithm. The classification algorithm used in the wrapper method of this study is the Support Vector Machine (SVM). SVM has been widely used for high-dimensional data classification.

The fitness value is the accuracy value of the correct prediction. In this study, the fitness value is used to select the features. Therefore, the value of SVM accuracy determines the fitness function of an individual. In this paper, we use one criterion fitness function containing accuracy for testing dataset as mentioned in (1):

$$fitness(x) = accuracy(x) \tag{1}$$

$Accuracy(x)$ is the accuracy of the SVM classifier trained using the feature subset of training data represented by x. Evaluation metrics are used to determine the effectiveness of the proposed proposal. Evaluation metrics use a confusion matrix. The actual output results are compared with the target output. There are four comparisons of evaluation metrics, namely True Positive (TP), False Positive (FP), True Negative (TN), and False

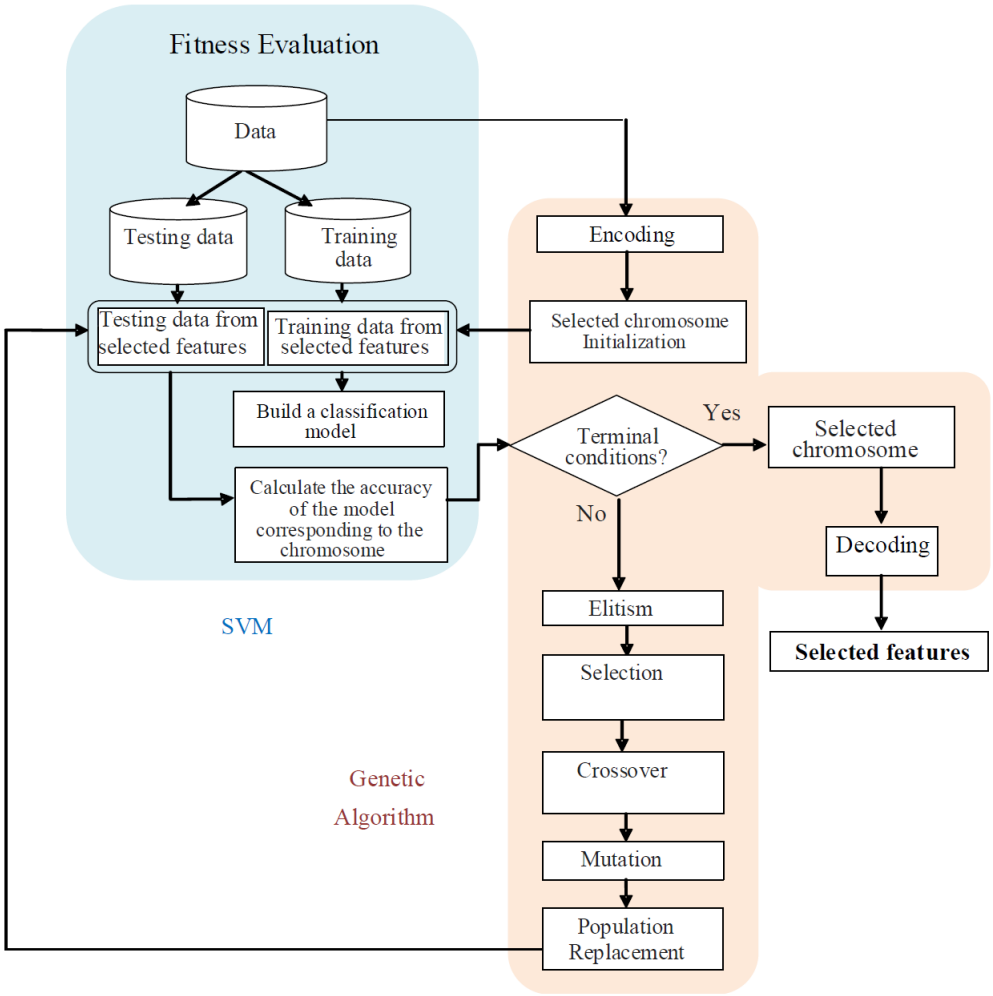

FIGURE 1. GASVM algorithm

Negative (FN) [8]. Below is the accuracy Equation (2):

$$accuracy(x) = \left(\frac{TP + TN}{Total\ number\ of\ instances}\right) * 100\% \tag{2}$$

Criteria restrictions are the maximum generation limit that has been fulfilled, taking the maximum fitness value – the terminal views from the criteria restrictions. MTS data used in this study have data training and data testing. The data testing is used to test its accuracy by using SVM.

Vectorization for Time of Observation Based on High Correlation (VecTOR). Features selection is made before the data mining process to ease the data mining process itself. In the introduction, it has been explained that the MTS matrix dimension will be quite significant. It is because the role of observation time affects the features or variables used. This simplifying process is emphasized in this research. This paper aims to form the simplest vectorization process, and the data are taken from the MTS set of data which is its observation time. This vectorization formation goes through several pre-processing stages. These stages are as follows.

a) There are N MTS with the different time series lengths, and the number of variables is m. An example of an MTS image can be seen in Figure 2(a).
b) The next step is to equalize the number of time series [9]. The uniformity of the number of time series is taken that has the most time series. MTS with the least time series is filled with null values so that no time series is truncated because there may be data

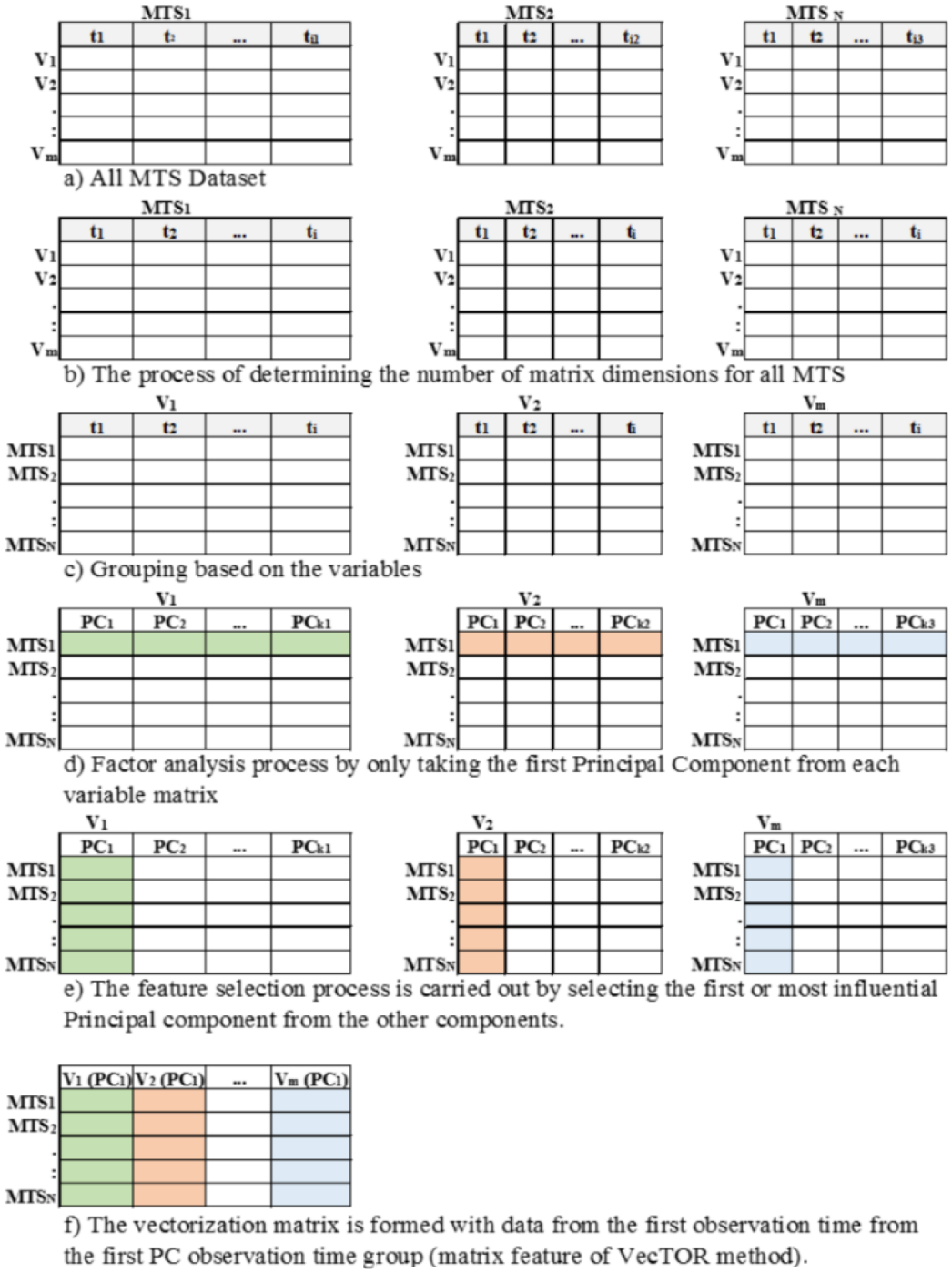

FIGURE 2. Proposed approach algorithm

that affect the time series patterns of other MTS data. Figure 2(b) shows value i is the number of time series for each MTS.

c) Each MTS data is grouped based on the number of variables. If there are N MTS data sets with m variables, there are m matrices with the observation time column (i). Figure 2(c) shows the grouping.

d) At this phase, the feature selection process is carried out twice. The first feature selection uses factor analysis, where this phase takes the most influential observation time from all observation times for each variable matrix. Feature selection here is a selection to choose the observation time using the largest loading of each Principal Component (PC) formed. Each matrix line is the selected observation time on each PC formed in each MTS. This process can be seen in Figure 2(d).

e) The second feature selection process is selecting the observation time only on the formed first Principal Component (PC_1). As a research development, the second stage was made. The first principal component of each variable is used to form a vectorization matrix. The formed vectorization matrix consists of PC_1 columns from each variable matrix. Figure 2(e) shows the selection of observation time from each variable matrix. Figure 2(f) shows the final result of forming a vectorization matrix called VecTOR (Vectorization for Time of Observation Based on High Correlation).

The vectorization formation is conducted by N row (time of observation) for each MTS with m column, which consists of the first principal component of each variable. This process is called Vectorization for Time of Observation Based on High Correlation (VecTOR). The $N \times m$ vectorization matrix is taken from the most influential observation time or one with the highest correlation in every matrix variable. Next, variables selection is made to the last vectorization matrix result using the GASVM wrapper method. The entire process of the proposed algorithm employs the filter wrapper method. First, the filter method is used to select the observation time feature of each MTS until the VecTOR vectorization matrix is formed. Then, the wrapper feature selection method selects variables from the VecTOR vectorization matrix. This VecTOR data matrix uses existing actual data, not the data extraction results. For example, suppose there is new data for testing its classification. Then, only one observation time is needed for the selected variable, unlike in [3,11], which requires further processing of MTS data or data extraction and then processed into a form adapted to the vectorization method.

3. Results and Discussion.

In applying vectorization formation variable selection, this study uses the GASVM algorithm. Selection strategy with a genetic algorithm uses Roulette Wheel, one-point crossover, and generation replacement. In SVM implementation, it uses a linear kernel function. The GA parameter is the mutation rate of 0.05, crossover rate of 100, the number of maximum generations is 100 and the population size is 50. Parameters for Naïve Bayes use Weka's default which is fold 5, seed 1, threshold 0.01.

3.1. Data and evaluation method.

This research uses CMU_S16_MOCAP data and the Wafer dataset. The selection of this dataset is because there are two classes for classification and represent MTS data with high and medium dimensions. This task predicts whether the subject is walking or running in CMU data. It contains information from 62 different combined positions recorded for various data totals [10]. CMU_S116_MOCAP has 62 variables. Each variable has an observation period of 127-580. While the training data is 29 data and the test data consists of 29 data. The Wafer dataset has 6 variables. The length of the time series is between 104-198. There are 298 training data and 896 test data. The position represents variable data until it is known as $V_1, V_2, \ldots, V_m$. Some features have the same value in the CMU data, namely feature (variable) 34 and feature 46. These variables are unable to be observed to find out their loading value or at the use of both phases. Besides, a variable cannot be observed yet due to its small value with a few digits below zero.

Those variables are variables 25, 26, 37, and 38. Therefore, these 6 variables are not included in the research. A wafer set of data is a group of time series data. Each file consists of a measurement order recorded from chamber vacuums sensory during the casting process of one silicon wafer that creates its semiconductor microelectronic [10]. Every thin bar or wafer consists of two types of normal and abnormal classifications. The abnormal wafer is shown by the distance of damage or problems that are casually found during the making process of semiconductors. Data consist of 6 sensory variables.

Validation of training data using cross-validation Kfold with $K = 5$ (CMU and Wafer data). Selection of K parameter is due to default weka application with $K = 5$. This

data set consists of different time series lengths so the first step is employed to unify the time series data dimension. In this study, we tested several wrapper algorithms, namely the Genetic Algorithm – Support Vector Machine (GASVM), Genetic Algorithm – Naïve Bayes (GABayes), the Forward-Naïve Bayes (FwdBayes) approach, and the Backward-Naïve Bayes (BckBayes) approach. In GABayes, a genetic algorithm is used to initialize chromosomes and fitness values as evaluators using Naïve Bayes. In FwdBayes, the forward approach is used to find a combination of variables and Naïve Bayes for the evaluator based on its accuracy level. The forward approach begins with one variable used and then two variables. In BckBayes, it uses a backward approach to choose a variable combination with Naïve Bayes as its evaluator. The backward approach's comparison begins from all variables, and the following variable is substituted.

3.2. **Performance of CMU dataset feature selection.** The proposed method in this research was tried out in a CMU set of data to know the classification performance based on its accuracy value. The classification according to variable or sensory will result from 56 variables matrix. Because variables 25, 26, 34, 37, 38, and 46 are not included (see the explanation in Subsection 3.1), the data total is 29. The time series length, between 127-580, requires unifying its matrix dimension. Therefore, this research utilizes maximum time-series length with an additional 0. Therefore, there is a matrix dimension of 29×580 in 56 variable matrices. It means that there are 29904 observation times in each variable. Using maximum time series is that we do not want to lose any observation time information from each data because this research is based on the correlation or the strong influence of observation time in each matrix variable.

As previously explained (Section 2), the feature selection process or filter method variable produces a vectorization matrix. Thus, the number of variables is 56. Then the feature or variable selection is carried out using the wrapper method.

Table 1 shows the number of the last selected variables. The total observation time in Table 1 is the number of seconds in the time series before feature selection. Feature selection is the lowest minimum total feature with the highest accuracy value after generating the selected variables and then further testing with test data. Table 2 shows the accuracy results with the test data and processing time. If there is no feature selection process, it will take a long time even though the accuracy is maximum. In Table 2, there is VecTOR-SVM. After the vectorization process, classification is carried out using the support vector machine without going through the wrapper method. From Table 2, it can be concluded, that VecTOR-GASVM has maximum accuracy and a faster time than without the feature selection process. On the other hand, VecTOR-FwdBayes and VecTOR-BckBayes have the least selected variables. However, the accuracy is less than the maximum compared to VecTOR-GASVM and the processing time is also longer. Table 2 shows features or variables selection in the formed vectorization. Feature selection is the lowest minimum total feature with the highest accuracy value. From the table, it can be seen that the method proposed has higher performance than the others that are with a total of 11 selected features and variables with 100% maximum accuracy. Therefore, predicting new data

Table 1. Results of selection of features or variables and their accuracy on CMU data

No	Method	Total observation time	Number of selected features	Accuracy (%)
1	VecTOR-GASVM	29904	11	100
2	VecTOR-FwdBayes	29904	3	96
3	VecTOR-GABayes	29904	15	93
4	VecTOR-BckBayes	29904	3	89

Table 2. Accuracy results of selected variables and access time on CMU data

No	Method	Number of selected features	**Accuracy (%)**	Time processing (seconds)
1	VecTOR-GASVM	11	**100**	0.56
2	No feature selection	29904	**100**	1.7
3	VecTOR-FwdBayes	3	**96**	2.5
4	VecTOR-GABayes	15	**93**	0.2
5	VecTOR-SVM	56	**93**	0.6
6	VecTOR-BckBayes	3	**89**	47.6

requires only 11 observation times from 11 variables used for the classification process from 56 variables with a length of time series between 127-580. Therefore, it proves that the results of research classification using the VecTOR-GASVM vectorization method are more effective and efficient without reclassifying when using data testing. The variables are variable 1 at t_{172}, variable 4 at t_{263}, variable 5 at t_{520}, variable 7 at t_{192}, variable 17 at t_{180}, variable 19 at t_{485}, variable 20 at t_{233}, variable 24 at t_{231}, variable 27 at t_{186}, variable 29 at t_{17} and variable 30 at t_{215}.

3.3. **Performance of Wafer dataset feature selection.** Similar to the usage of the CMU dataset, the Wafer dataset will use several feature selection methods to know the best accuracy. The accuracy result and selected observation time can be seen in Table 3. FS is the total selected variable features. Like CMU data, Wafer data's highest accuracy value is obtained using the proposed method, features selection GASVM. The maximum accuracy level obtained is 2 variables. It means that it takes only 2 observation times from 2 variables which are only variable one at t_{128} and variable two at t_{95}. Therefore, the computational time will be more effective and efficient, with a smaller data size. Table 4 shows the processing time of the selected feature or variable. VecTOR-GASVM looks not have the fastest process but has the highest accuracy.

Table 3. Results of selection of features or variables and their accuracy on Wafer dataset

No	Method	Total observation time	Number of selected features	Accuracy (%)
1	VecTOR-GASVM	954	2	97.98
2	VecTOR-SVM	954	6	89.93
3	VecTOR-GABayes	954	1	68.45
4	VecTOR-FwdBayes	954	6	68.12
5	VecTOR-BckBayes	954	6	68.12

Table 4. Accuracy results of selected variables and access time on Wafer dataset

No	Method	Number of selected features	**Accuracy (%)**	Time processing (seconds)
1	VecTOR-GASVM	2	**97.98**	0.79
2	No feature selection	954	**95.30**	0.75
3	VecTOR-SVM	6	**89.93**	1.2
4	VecTOR-GABayes	1	**68.45**	0
5	VecTOR-FwdBayes	6	**68.12**	0.23
6	VecTOR-BckBayes	6	**68.12**	0.52

4. **Conclusions.** This research gives a new method for forming vectorization for the multivariate time series data. The vectorization process is done by finding the highest loading value from the first principal component. This method is called Vectorization for Time Observation Based on High Correlation (VecTOR). The final result is a vectorized column of each variable at a specific observation time, and the row is the recorded data. In this research, the data used for vectorization is original observation data until the end of the feature selection process. It will ease the process if new data is classified, for example, to know its prediction label. Unlike previous studies [11], this research uses extracted data. This research uses the filter wrapper method. The filter method is carried out during the observation time selection process for the vectorization matrix. The wrapper method is used for variable selection from vectorized matrices. The selection of the best features is based on the highest accuracy with the least number of selected variables. In CMU data, the VecTOR-GASVM method has an accuracy of 100 per cent with 11 features selected. It means that it takes only 11 observation times from 11 variables to get a 100% accuracy value. For the Wafer set of data, VecTOR-GASVM has an accuracy of 97.98 per cent with 2 features selected. It takes only 2 observation times from 2 variables to get a 97.98% accuracy value.

The formation result of vectorization column and row can be examined using other methods employed in this research, especially soft-computing based methods, to know the higher accuracy level with the least total of selected features variable. However, other experiments are needed to confirm this conclusion using other MTS datasets with different dimensions.

REFERENCES

[1] J. J. A. Mendes Junior, M. L. B. Freitas, H. V. Siqueira, A. E. Lazzaretti, S. F. Pichorim and S. L. Stevan, Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by sEMG in eight channels armband approach, *Biomed. Signal Process. Control*, vol.59, doi: 10.1016/j.bspc.2020.101920, 2020.

[2] T. N. Lal et al., Support vector channel selection in BCI, *IEEE Trans. Biomed. Eng.*, vol.51, no.6, pp.1003-1010, 2004.

[3] K. Yang, H. Yoon and C. Shahabi, A supervised feature subset selection technique for multivariate time series, *FSDM*, 2005.

[4] A. D. Rahajoe, E. Winarko and S. Guritno, A hybrid method for multivariate time series feature selection, *Int. J. Comput. Sci. Netw. Secur.*, vol.17, no.3, pp.103-111, 2017.

[5] R. Kohavi and G. H. John, Wrappers for feature subset selection, *Artif. Intell.*, vol.97, pp.273-324, 1997.

[6] E. Alba, L. Jourdan and E. Talbi, Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, *IEEE Congr. Evol. Comput.*, vol.7, pp.284-290, 2007.

[7] T. Zhou, H. Lu, W. Wang and X. Yong, GA-SVM based feature selection and parameter optimization in hospitalization expense modeling, *Appl. Soft Comput. J.*, vol.75, pp.323-332, doi: 10.1016/j.asoc.2018.11.001, 2019.

[8] R. Dash, A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study, *J. King Saud Univ. – Comput. Inf. Sci.*, vol.32, no.2, pp.232-247, doi: 10.1016/j.jksuci.2017.08.005, 2020.

[9] T. Górecki and M. Łuczak, Multivariate time series classification with parametric derivative dynamic time warping, *Expert Syst. Appl.*, vol.42, no.5, pp.2305-2312, doi: 10.1016/j.eswa.2014.11.007, 2015.

[10] R. T. Olszewski, *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*, Ph.D. Thesis, Carnegie Mellon University, 2001.

[11] H. Yoon and C. Shahabi, Feature subset selection on multivariate time series with extremely large spatial features, *The 6th IEEE International Conference on Data Mining – Workshops (ICDMW'06)*, Hong Kong, China, 2006.