ISSN: 0974-8571 Vol. 13 No. 1 June, 2021

International Journal of Computational Intelligence in Control

Classification Algorithms and Feature Selection Techniques for a Hybrid Diabetes Detection System

Bassam Abdo Al-Hameli
IBM Centre of Excellence, Faculty of Computing, Universiti Malaysia Pahang, Pahang, 26600, Malaysia
AbdulRahman A. Alsewari
[1]IBM Centre of Excellence, Faculty of Computing, Universiti Malaysia Pahang, Pahang, 26600, Malaysia
Abdulaziz Saleh Alraddadi
[2]College of Computer Science and Engineering. Taibah University Yanbu. Saudi Arabia (alraddadi1@yahoo.com)
Arafat Aldhaqm
[3]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Johor, Malaysia (email: mrarafat1@utm.my)

Abstract: The concept of combining classifiers with feature selection techniques recently became a new trend for improving the performance of classification algorithms in the light of the explosive growth of data, which healthcare data classification has become a very difficult task. Early detection systems for diseases have been used to help detect harmful activities in the human body, including diabetes, which has become the main cause and origin of many diseases. Thus, several researchers have proposed Diseases Detection Systems (DDSs) with a combination of approaches such as Machine Learning methods and algorithms inspired by nature to deal with the difficulties of diseases detection problems. Diabetes Detection System (DiDS) is proposed to detect early factors of diabetics. In this study, optimization of Hidden Naive Bayes (HNB) and Naive Bayes (NB), and Decision Tree (DT) binary algorithm supported with discretization within Feature Selection (FS) techniques were combined to reduce the data dimensions in order to produce better classification performance and accuracy in most cases that use less time for training as well. To evaluate the performance of proposal system, Pima Indian Diabetes (PID) dataset has been used. The proposed system analysis was done using the performance measures Sensitivity, Specificity, and Receiver Operating Characteristics (ROC) curve. The experiential results analysis shows that the HNB classifier improves the performance of DiDS in terms of accuracy and predicting states.

Index Terms - HNB; Diabetes Detection System; PID dataset; Feature Selection

1 Introduction

Machine learning algorithms use sample data to build a mathematical model. Supervised and unsupervised programming are the two types of machine learning activities. In supervised learning, training data is classified, and response variables may be discrete/qualitative (for the classification task) or continuous/quantitative, as in the case of this study, for machine learning. Since the dependent variable can be separated into binary or multiple classifications, Machine learning is a sorting activity for healthcare diagnostics. Machine learning algorithms will only make correct predictions if the testing data sets are free of harmful data. Several machine learning algorithms that have been shown to provide good healthcare diagnostics are discussed below. In the healthcare industry, machine learning algorithms aid in the prediction of preferences [1]. In the healthcare industry, machine learning algorithms aid in the prediction of preferences. By reviewing various AI algorithms either alone or in combination with other methods, and has been successfully deployed in medical diagnostics such as stroke, blindness, heart attack , cardiac disease, kidney failure , Amputation, etc.; Which was the first reason for their existence is diabetes [2].diabetes can lead to complications in many parts of the body and increase the risk of dying prematurely [3], [4] The existing state of technologies and data sets for healthcare providers, who are already struggling to streamline data from legacy systems.

Especially in the last few years, it has been observed that high-dimensional data is increasing exponentially. The discovery of diabetics patients faced many problems that dealt with the ability to discover patterns and features in the patient data system. Since we need high dimensional data to identify the infected patients. Disease prediction plays an important role in data mining; disease detection also requires several tests and clinical examinations on the patient. However, the use of data mining techniques and machine learning algorithms in the field of healthcare data classification can reduce the number of tests. This reduced test suite plays an important role in performance and time. Extracting and selecting features in healthcare data is important because it allows clinicians to know which features are most important for diagnosis, such as age, weight, symptoms, etc.[5], [6], and this will help clinicians

diagnose disease more efficiently and detect cases earlier. To solve this problem, feature selection methods were investigated to determine the original data for the best subgroups for classification accuracy. Detection Systems (DSs) are common software systems that automate monitoring of the events occurring in database systems or data warehouses, analyzing them for influenced factors or signs of specific data or attributes to assist in determining certain problems or different patterns. Detection systems are multiple and oriented systems for Learning and Controlling Behavior and Reflexes and Instrumental Conditioning, Habits, and Goal-Directed Actions, which are employed in several areas, including detecting diseases, intrusion, failure, and fault problems in software and hardware [7], [8]. Diseases Detection Systems (DSSs) are used to plant care [9] and healthcare [10] with the specification of the kind of disease (e.g., heart and cancer diseases). Intrusion detection systems are widely used in the domains of research and software and to detect anomalies to catch intruders and hackers due to increasing network attacks and cyber-attacks [11], [12].

.

The study aims to conduct an accurate statistical analysis to investigate Pima Indian Diabetes (PID) dataset characteristics and invest features that have impact factors for improving classification performance, FS techniques are used to decrease overlapping data and special features for increased performance and accuracy. This study appears conceptual framework to FS techniques roles on benchmark datasets and their impact on associated components and selects the most relevant features. This paper will contribute to finding a better subset of features to represent the dataset that can be accessed globally, using feature selection techniques including BPSO, GR, GI, IG, ReliefF, and others with NB, HNB, and DT classification algorithms.

2 Material and Method

2.1 Classification Algorithms

The Classification Problem involves data, which is divided into two or more groups, or classes. Classification is a classic data mining task and has roots in machine learning. Classification is one field of machine learning supervised learning and predictive analysis, i.e. (allocates data into separate categories. the most popular classification algorithms used in healthcare such as Random Forest, Support Vector, Naïve Bayes, Nearest Neighbor, Decision Trees, and HNB.

This paper utilizes a hybrid approach to evaluate our proposed system with feature selection methods for PID dataset classification. Therefore, we chose 8 feature selection techniques, one hybrid BPOS classifier with NB based feature selection, and two evaluation performance measures for features (instances) are cut-point and AUROC measures, three classification algorithms based on PRE-FS and POST-FS employed, 8 metrics of the binary classification evaluation, and two functional methods to evaluate efficiency, effectiveness and stability of the model and for validating each evaluation approach.

2.1.1 HNB

HNB proposed by Jiang et al. [13] uses a discrete structural model and hence requires the discretization for preprocessing with continuous signal attributes, such as expression microarray data. It is used to relax the conditional independence assumption of Naïve Bayes classifier (NB). HNB classifier solves the problem of conditional independence between features by building a hidden parent for each value [14]. HNB classifier seeks to improve performance using approaches that rely on finding correlations amongst the features, reducing the strong assumption of independence that HNB is based on [13].

$E(a_1, a_2, \ldots., n)$, $Each\ attribute\ Ai\ has\ a\ hidden\ parent\ a_{hpi}$

$$c(E) = argmax\ P(c) \prod_{i=1}^{n} \quad P\left(a_{hpi}, c\right),\ E = a_i \quad (1)$$

2.1.2 NB

NB classifier is a straightforward probabilistic classifier based on the Bayes theorem and strong "naïve" independence conventions. The Naive Bayes classifier was proposed in 2006 to phishing email filtering in Microsoft. Naive Bayes assumes that human characteristics are conditionally independent of one another [15].

Significantly recently, the use of the Naive Bayes approach increased, as a result of the increase in data analysis, many researchers such as Jeffrey's de Finette, Savage, and Lindley have developed Bayesian Analysis. According to the subjective model, the probability is the self-uncertainty of the observer. The probability value is emotional, and the new evidence can be changed upon arrival [16].

$$P(B) = \frac{P(B|A) * P(B)}{P(A)} \quad (2)$$

P (A) = Independence probability of event A, P(B) = Independence probability of event B, P(B|A) = the probability of event B when event A is known ("Likelihood, Conditional Probability") P (A|B) = the probability of event A when event B is known ("Posterior, Posterior Probability"). The Naive Bayes classification is one of the probability methods that predict the instances of the class that has the highest probability and it works as flows. Assume that A and B to be random events.

Hidden Naïve Bayes and Naive Bayes are probabilistic machine learning algorithms based on the Bayes Theorem, used in a wide variety of classification tasks.

2.1.3 Decision Trees

Decision Trees specify the sequence of decisions that need to be made and the resulting recommendation, which naturally leads to a style of representation called a decision tree. Decision trees are an estimation technique used in classification, grouping, and prediction models and subdivide the research area related to a problem into subgroups.

Tree of Decisions (DT) Nonparametric classifiers are simple to understand and imagine, and they can easily capture nonlinear patterns. ID3, C4.5 (ID3 extension), and decision tree CART formats are among the formats available. J48 is a decision tree learner that is created by

C4.5 and implemented in Weka. It can be used to describe the function correctly, but DT can increase the amount of irritating data included, which can be minimized by using ensembles [13].

2.2 Feature Selection (FS).

Adding or removing Not influential features is main purpose of feature engineering. Feature selection is playing a significant role in any classification task ,and essential for simpler, faster, more reliable and robust ML models. The goal is to maintain accuracy and stability, improve operating time and avoid overfitting , get rid of uninformative features, simplify the model, as well as reduce noise in data and confront multicollinearity. Feature selection technology distinguishes redundant or unrelated data which can be removed without losing too much information, as well as how to select a subset of potential attributes or variables from a dataset. Filter-dependent algorithms are filter dependent, Embedded, Wrapper Based, and hybrid [14].

Studies [17], [18] concluded on the PID dataset in which the researcher used the following five algorithms and compared them with a choice of pre-and post-features selection. Support Vector Machine, Regression, Bayes Net, Naïve Bayes, Decision Tree are chosen. The classification accuracy results with slight differences in pre-employ features selection techniques were 77.47, 77.34, 78.25, 76.30, and 77.6. The results of post-employed features selection were, respectively, 77.73, 77.60, 78.25, 77.60, and 79.81. In another Research paper by Mukherjee et al. [19], Naive Bayes suggested three common feature selection methods to reduce data sets. The methods are information Gain, Gain Ratio, and Correlation-based feature selection. Selected features set gave effective efficiency with more accurate results than the other methods. Among the most prominent of these technologies are the following:

*2.2.1 Information Gain (IG***)**

The Information Gain (IG) method of feature selection is fairly simple to use. When a particular feature is used to group values of another (class) feature [20], IG measures the amount of information in bits about the class prediction and the decrease in entropy. The entropy of Y is indicated by the letter H(Y) , where $P(x_i/y_i)$ is the conditional probability of xi given yi.

$$H(Y) = -\sum_i \quad P(y_i) \sum_i \quad P(x_i/y_i)\, log_2\,(P(x_i|y_i)) \quad (3)$$

,and The entropy of a feature X is defined as

$$Info\ D = H(X) = -\sum_i \quad P(x_i) log_2(P(x_i)) \quad (4)$$

It is denoted as IG(Y|X).

$$IG(Y) = H(X) - H(Y) = H(Y) - H(Y|X) \quad (5)$$

According to the equation IG(Y|X), we can conclude that the larger the IG value, the greater the impact of the corresponding features vector for the prediction methylation site. The decision tree (DT) has two types of entropy, Entropy of one attribute and two attributes. The information gain is related to a reduction in entropy after a dataset is split on a feature. Building a decision tree is about obtaining attribute that returns the highest information gain.

2.2.2 Gain Ratio (GR)

GR is a modification of information gain (IG) that addresses the issue of bias toward features with a large set of values that occurred in IG. GR is used as a splitting measure to select the most discriminative feature at each step of the classification process during training the model [21]. When GR chooses an attribute, it considers the size and quantity of values. When all data belongs to one branch attribute, GR should be small, and when data is evenly distributed, it should be big.

$$SplitInfo(Y) = \sum_{i=1}^{n} \quad \frac{|(Yi)|}{|(Y)|} * log_{\,2}\,(\frac{|(Yi)|}{|(Y)|}) \quad (6)$$

$$GR = \frac{IG(Y)}{SplitInfo(X|Y)} = \frac{IG}{H(X)} \quad (7)$$

GR values always fall in the range [0, 1]. A GR = 0 indicates no relation between Y and X, and the value of GR = 1 indicates that the knowledge of X completely predicts Y.

2.2.3 Relief-F (RF)

Kira and Rendell [22] proposed the Relief method. The original Relief method based feature quality on how effectively their values differentiate between instances close to each other. Relief selects a random example from the data and then finds its nearest neighbor from the same class and its opposite class. ReliefF's main notion is to estimate features based on how effectively their values discriminate between examples that are close to each other. More information is available in [23]. Where Relief looks for two of its nearest neighbors: one from the same class (called nearest hit, or "H") and the other from a different class (called nearest miss, or "M"). Relief's original approach picks m training examples Ri, I = 1,,...,m at random, where m is a user-defined parameter, and the weight of attribute A is determined as follows:

$$W[A] = W[A] - \frac{1}{m} \sum_i \quad diff(A, R_i, H) + \frac{1}{m} \sum_i \quad diff(A, R_i, M) \quad (8)$$

Function diff (A, R, V) calculates the difference between the values of the attribute A for two instances R and V.

2.2.4 Gini Index (GI):

There are several possible methods and adaptations for Information Gain. One of the most prominent alternatives is the Gini Index [24], which is an impurity measure of D rather than an entropy metric. GI is a supervised technique with a simpler computation than IG, and it was first employed in decision-tree algorithms.

$$GI(D) = 1 - \sum_{i=1}^{m} \quad P_i^2 \ , \quad (9)$$

where D is dataset, Pi is the probability frequency of class i (part training instances belonging to class i)

2.2.5 Correlation-based Feature Selection (CFS) :

The CFS method was suggested by M. A. Hall [25]. The technique identifies factors that are both substantially associated with the final prediction and poorly associated among themselves. Considers the individual predictive capacity of each feature and the degree of redundancy between them when determining the value of a subset of

characteristics. Subsets of elements with low intercorrelation but strong correlation with the class are selected. Duplicate features are also filtered away due to their significant correlation with the remaining features.

$$M_s = \frac{f * r_{cf}}{\sqrt{f + f(f-1) r_{ff}}} \quad (10)$$

where the merit (M) heuristic is calculated for the containing f features in the dataset and M_s is the merit of a subset (s) of features, and it's the correlation between the summed components and the outside variable, r_{cf} is the mean feature class correlation, and r_{ff} is the average of feature class inter-correlation.

2.2.6 Analysis of Variance (ANOVA)

ANOVA is a parametric statistical test technique used to assess the importance of features, in other words, it's (in part) a significance test. It uses a null hypothesis (nonparametric tests) and ranks all the entropic features, which attempt to express one dependent variable as a linear combination of other features or measurements. The test has been carried out with a high percentage confidence interval, 0.05 significant level, and linear polynomial contrast [26].

$$F = \frac{variance\ (differences)\ between\ sample\ means}{variance\ (differenc\ es)\ expected\ with\ no\ treatment\ effect} \quad (11)$$

2.2.7 Fast Correlation Based Filter (FCBF) algorithm

In both the preparation and preprocessing parts, the FCBF method decreases the dimensionality of the dataset [27]. The selection of an appropriate classification algorithm is also critical. FCBF begins by identifying a set of features F that are highly linked to the class but not to other features. Correlation between two features are measured. Relevant features are selected goodness of feature for classification from the original dataset such that it is highly correlated to any other class. There are two ways to choose FCBF features; Choose one feature that is important to the class and remove other characteristics iteratively depending on the chosen feature.

2.2.8 Cut-point method.

The word "cut-point" denotes to a real value inside a range of continuous values that splits the range into two intervals, one less than or equal to the cut-point and the other greater. Split-point is another name for cut-point. Class information is utilized to determine the correct intervals induced by cut-points using supervised discretization algorithms.

2.2.9 Binary Particle Swarm Optimization (BPSO)

The method was proposed by Kennedy and Eberhart to allow PSO to operate in the space of binary problem and search [28]. BPSO is a new feature selection method for unsupervised learning and is a combined approach, presenting a new neighbor selection strategy to identify salient features, where POS performance is limited due to select features [29].

Binary PSO has been implemented in many areas, including feature selection, although it has not been thoroughly investigated. The overall goal of Binary PSO in this thesis is to select the feature to achieve good rating performance.

$$x_{ij}\ (t+1) = x_{ij}\ (t) + v_{ij}\ (t+1) \quad (12)$$

$$x_{ij}(t+1) = \{1,\ rnd < sig(vid(t+1))\ 0,\ otherwise \quad (13)$$

Then x_{ij}=1 else x_{ij} =0 if x_{ij}=1,2,…,n, d feature is chooses.

Lin and Yu [30] suggested Particle Swarm Optimization algorithm (PSO) using the Naive Bayes weighted method that uses PSO as a research function in simultaneously maintaining the integrity of each feature of the datasets. Their goal was to achieve improved classification accuracy while avoiding the loss of information as a standard for the experiment used for UCI datasets. Sengottuvelan, P., & GopalaKrishnan proposed Hybrid PSO with Naive Bayes. The aim was to analyze the trait that causes a decrease in the accuracy of predictive analysis of disease incidence, Expected results achieved and reached 96.6% [31]. Other researchers provided a combination of PSO-Naive Bayes for 19 datasets, and the proposed approach had a classification accuracy of 84.63% [32].

2.2.10 Fitness function

It is used to assess the competence (performance quality) of candidate solutions. The selection of fitness is an important aspect in the classification algorithms. The precision value is the most important consideration when creating a fitness function, P in [0,1] [33].

$$Fitness = (P), \quad (14)$$

$$when\ precision\ (P) = \frac{TP}{TP+FP}, \quad (15)$$

whereas FP indicates value of the False Positive and TP indicates to the True Positive.

2.2.11 Stability Measure (SM):

The robustness of a FS techniques and algorithms are measured by their stability. Stability indicates the FS applied to different subsets produces stable output, as well as it requires defining a similarity measure that assesses the commonality of a pair of feature subsets. Robustness necessitates the stability of the chosen features. The final classified model may be degraded due to Unstable feature selection performance, which leads to failure to identify the most relevant features [24]. When the training set modifies, the stability is calculated as follows:

$$SM = \frac{2}{R(R-1)} \sum_{i=1}^{R-1} \sum_{j=i+1}^{R} \frac{|Si \cap Sj|}{|Si \cup Sj|} \quad (16)$$

Where S_i and S_j are two feature sets selected (Fs) by each FS method from dataset (D), which measures the similarity between S_i and S_j , R is total of feature subsets.

The strength of the feature subset obtained from multiple training sets of the same distribution is characterized as stability. The approach is deemed stable when the parameter of the selection technique causes modest changes to a feature subset. Due to a failure to pick the most relevant features, unstable feature selection performance affects performance in the final classifier [24]. Different metrics can be used to assess robustness. The following [34] distinguishes these measures for evaluating the stability of feature selection methods: Feature-focused vs. subset-

focused: the former assesses a feature selection technique by combining all feature subsets, whilst the latter analyses feature similarities in each pair of two subsets [35].

3 Diabetes Detection System (DiDS)

3.1 PID Dataset Descriptions

Machine learning in healthcare is one of the most complex industries and the most interest by researchers. Therefore, PID dataset is one of common and important dataset in healthcare systems. We all know that diabetes is one of the most common dangerous diseases. You can use this dataset in your diabetes detection system. This dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of this dataset is to predict whether or not a patient has diabetes based on specific diagnostic measurement.

3.2 Data Preprocessing

Pima Indians diabetes (PID) dataset consists of D= 8 numerical medical attributes and c = 2 classes (tested positive "affected by diabetes" or negative for diabetes "non-diabetic"). There are n = 768 instances which divided into 500 negative instances and 268 instances of positive .

With data preparing, all data will be changed in value into discrete bins form. The attributes descriptions are shown in Tab.1 below

.**Table 1**: description of Parameters and features in PID Datasets

Sub No.	Abbre.	Type	Mean	Stand. Dev.	Min	Max	Distinct	Unique	Missing Value
S1	Pregnant	Numeric	3.8	3.4	0	17	17	2 (0%)	No
S2	Plasma Glucose	Numeric	120.9	32.0	0	199	136	19 (2%)	No
S3	Blood Pressure	Numeric	69.1	19.4	0	122	47	8 (1%)	No
S4	Skin	Numeric	20.5	16.0	0	99	51	5 (1%)	No
S5	Insulin	Numeric	79.8	115.2	0	846	186	93 (12%)	No
S6	Mass	Numeric	23.0	7.9	0	67.1	248	76 (10%)	No
S7	Pedigree	Numeric	0.5	0.3	0.078	2.42	517	34 (45%)	No
S8	Age	Numeric	33.2	11.8	21	81	52	5 (1%)	No
S9	Class	Nominal	-	-	0	1	2	0	No

3.2.1 Discretization

The discretization approach performs better when there is a significant amount of training data, due to it learn to suit the data distribution. Discretion is often selected over the distribution technique since Naive Bayes is typically employed when a significant quantity of data is available (because it is computationally more expensive and models can generally attain higher accuracy).It is important to convert the continuous attribute to discrete to ensure the efficiency of the system ,distributes into selected number of bins equally , and to solve the problem of appear new value and to increase speed and ensure the effectiveness of the system and also to solve the problem of NB classifier when new value appeared in test dataset that didn't appear in learning phase.

Researchers are realizing that in order to achieve successful feature selection is an indispensable component [29]. The feature selection approach takes enormous amount of time to find minimal subset of features. The new researches in this area focus on reducing training and testing set time for the purpose of efficient research. Hence this work proposes an effective and efficient approach to find the best subsets, which is compared with others methods. The methodology for proposed system named diabetes detection system (DiDS) adopted in this work for the diagnosis of disease is shown in Fig. 1 and : General structure in fig 2

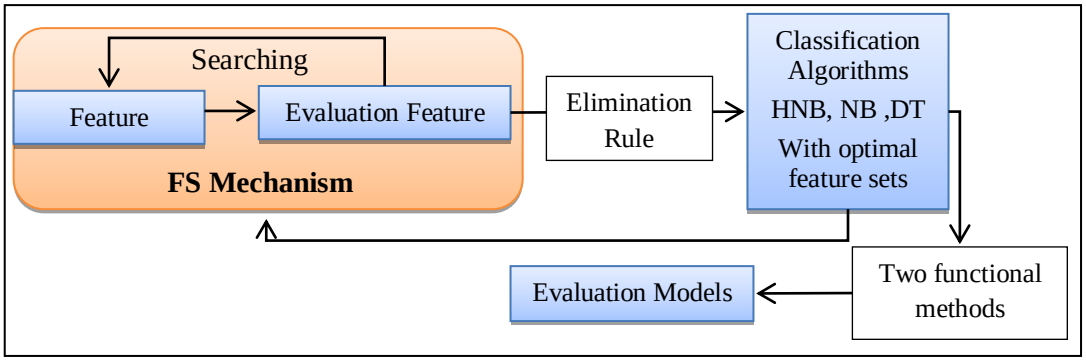

Figure 1: Block diagram of the proposed system (DiDS).

Our proposed system divided into two phrases. In the first phrase, for building a set of PID dataset to detect ideal features , we conducted 11 experimental runs for feature selection techniques and algorithms and evaluation sub sets of features selected by ranking method of FS techniques , and finally elimination rule. In the second phrase, we used two types of 10-fold cross-validation and percentage split techniques for calculating stability and other metrics. Both feature selection and classifier training depended on the training set. We estimate run time by fitness function,

stability, TPR, TNR, accuracy, AUROC and precision, Sensitivity, Specificity, F-score over 36 runs.

Input: PID classification dataset
Output: (S best) with rankings of FS techniques and algorithms
Procedures:
1- PID dataset is preprocessed, operations are performed on PID dataset as below
2- Compute distance of dataset (move to FS methods)
3- Discretization
4- Feature selection methods is using to remove sets of characteristics from the processed dataset.
a. Find size of training set D_{size} and probability each class (c).
b. Compute the entropy of each class (c) to find info D to M_i by use Eq. (5,6,7,9)
c. Compute data for all features ($f_i \in F$) and target class (c) as a score and compute frequency of each value in training set dataset calculated by Eq. (5,7,8,9,10,11,13).
d. Choose sets of features with the largest rating $argmax\ f_i \in F\ (M(f_i, c))$ and insert it to the subsets of selected features ($S_{selected}$)
e. Compute the possible rating based on the data in ($f_i \in F$).
f. Repeat step b and c until a sufficient number of features are selected ($S_{selected}$).
g. Apply Elimination Rule.
5- The final processed dataset is uploaded for training and testing sets
6- Evaluation selected features ($S_{selected}$).
Input: $S(f_1, f_2, \dots, f_n, c)$ as training dataset of optimal subset of features ($f_i \in$ Sselected)
Output: HNB, NB, DT classifier predictive model with FS techniques
7- The DT, NB, HNB algorithms are employed in training dataset by use Eq. (1,2) for DiDS.
8- Cross-validation and percentage split test techniques are used for model creation in the algorithms
9- Computing Accuracy and Performance metrics by (15,18,19,20)
10- Results analysis and Evaluation the proposed system by suitability measure by use Eq. (15,16).
11- Computing fitness value,
12- Determine and evaluate the best fitness value, Compare the Pi of every particle's fitness value, If the current fitness value is better assigned value to Pi, unless keep Pi value (pre value).
13- Output the global optimum

Figure 2: General structure of the proposed system (DiDS).

The FS techniques are faster, scalable, algorithm independent and great computational compare to other techniques. some of techniques elect the m subset features from the original n features which maintain the relevant information as in the whole feature set, knowing that m<n. For this ,it's possibility to create M_i from subsets features , which is a partial sets of the n; where $M_i \subset n$ and m $\in M_i$.

4 Results and Discussion

Classifiers based features selection techniques are approaches that work based on building full models and create new expectations with many adaptive methods to produce the final results. classifiers based feature selection for selecting the optimal feature set that enhances the predictive accuracy of the classifier. The features are ranked using feature relevance and the feature subset is evaluated by applying to Classifiers to select the optimum one which produces better predictive accuracy.

Initially, we consider the training data set (D), which consists of S number of traits and N number of data. Here, a number of features are given to the estimation function in order to transfer the input data to separate data and each row comprises N features; $0 < n \leq 8$. The main property of estimating is changing the data value to a specified interval, which means changing the range of data value to a specific period of time. The estimation process by using cut-point method is explained below.

Table 2: obtained results for the PID dataset Using Cut-point method

Features	Cut -point	No of points
Preg	0 to 6.5 ,6.5 to 17	2
Plas	0 to99.5,99.5 to 127.5,127.5 to154.5, 154.5 to 199	4
Pres	All	1
Skin	All	1
Insulin	0 to 14.5,14.5 to 121, 121 to 846	3
Mass	0 to27.85,27.85 to 67.1	2
Pedigree	0.08 to 0.53,0.53 to 2.42	2
Age	21 to28. 5,28.5 to 81	2
Class	0,1	2
Total		19

By calculating the number of periods cut-points in the data unit, through which points are determined, and this shows the diversity of data in each feature and so that it removes the redundant data and features. Tab.2 shows the features that obtained the highest cut-points, which it's possible to affect the performance of the classifiers in the classification state.

Algorithms and techniques of feature selection (Fi) measure the expected reduction in uncertainty associated with a random feature and resulting in (Fs). The applied feature selection techniques for picking best subsets are: IG, GI, GR, ReliefF, CFS, FCBF, ANOVA and BPSO_NB. In Tab.3, we observe that IG, GI, GR are sharing the order of

features and the subsets of useful (more discriminating) and useless features ($Fs \subset Fn$) for passing to training and testing set stage, where three features emerged in the bottom of the list as a useless feature (skin, Pedigree, and BPressure). Whereas, the ANOVA test method differed from others in the results, as it showed in the ranking of the features at the bottom of the list and they were close to zero, while the other features had a high ranking (skin and BPressure). ReliefF and FCBF emerged zero values for feature that consider useless features. Tab.3 shows entropy of FS analysis as the maximum and minimum values of each feature. Otherwise, correlation-based feature selection (CFS) offers a heuristic of individual features for predicting the class labels, as shown in the Tab.4.

Table 3: FS techniques used on PID dataset

	#	Info. gain	Gain ratio	Gini	ANOVA	ReliefF	FCBF
Plasma glucose		0.170	0.085	0.101	213.162	0.036	0.131
age		0.081	0.041	0.048	46.141	0.009	0.059
mass		0.079	0.039	0.044	71.772	0.010	0.057
insulin		0.055	0.030	0.031	13.281	0.000	0.000
pregnant		0.043	0.021	0.028	39.670	0.010	0.000
thickness		0.036	0.018	0.022	4.304	0.014	0.000
pedigree		0.022	0.011	0.014	23.871	0.013	0.015
blood pressure		0.015	0.007	0.009	3.257	0.009	0.000

Table 4: correlation based FS among features before selection from PID dataset

Features	Preg	Mass	Insu	Plas	Skin	Pedi	Pres	Age	Class	Ranked
Preg	-	-	-	-	-0.08	-	-	-	-	0.24
Mass	0.02	-	-	-	0.39	0.14	-	-	-	0.30
Insu	-0.07	0.20	-	-	0.44	0.19	-	-	-	0.14
Plas	0.13	0.22	0.33	-	0.06	0.14	0.15	0.26	-	0.23
Skin	-	-	-	-	-	-	-	-	-	0
Pedi	-0.03	-	-	-	0.18	-	-	-	-	0.17
Pres	0.14	0.28	0.09	-	0.21	0.04	-	-	-	0
Age	0.54	0.04	-0.04	-	-0.11	0.03	0.24	-	-	0.31
Class	-0.04	0.01	0.00	0.01	0.02	-0.04	0.02	0.01	-	0

The performance of features is visualized by using a Receiver Operating Characteristic (ROC) curve. The Area Under the ROC curve (AUC) is used as a tool for comparing the performance of the features on the removal of each feature. each feature has curve that has a larger AUC is better than the one that has a smaller AUC.

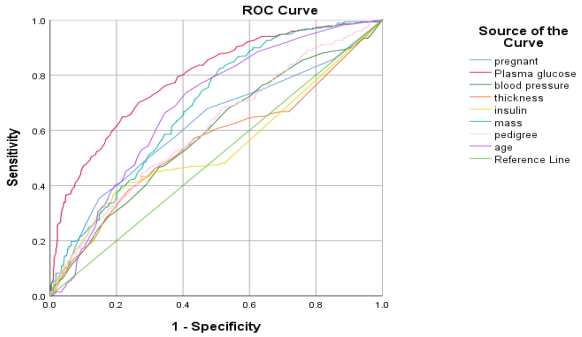

Figure 3: ROC curve analysis for each feature from PID dataset

At first glance, features of the PID dataset found that there are sets of features distinguished from each other. ROC and AUC curve analysis indicates that the top-classified features appear above the reference line, they are the optimal sets of features that will give the most accurate classification model. from these features respectively the Plasma, mass, age, BPressure, and Pedigree, which show as the best features that support classification algorithms. However, features emergence unusual features that placed under this line such as thickness(skin), Insulin, and BPressure, these classify as least effect and get the least accurate classification model.

The Area Under ROC (AUROC) for the best sets of features are (0.788, 0.688 0.687, 0.62, 0.606); respectively (Plasma, Mass, Age, Pregnant, and Pedigree); and for the non-optimal feature subset it is 0.586, 0.554, 0.538; respectively (blood pressure, skin, and insulin). according to Eq. (17), results of the AUROC analysis show the minimum and maximum values of each feature as in the following Tab.5.

$$AUC = \int_0^1 ROC(t)\, dt. \quad (17)$$

Table 5: AUROC analysis for each feature from PID dataset

Test Result Variable(s)	S1	S2	S3	S4	S5	S6	S7	S8
Area	0.62	0.788	0.586	0.554	0.538	0.688	0.606	0.687

Whereas t = 1 - Specificity and ROC(T) is sensitivity.

The test result variable (Pregnant, Plasma, BPressure, Skin, Insulin, Mass, Pedigree, Age) has at least one tie between the positive actual state group and the negative actual state group. Statistics analysis may be biased.

BPSO algorithm lately proven efficiency to select the optimal features of dataset, therefore, we proposed support the BPSO algorithm with NB for getting the ideal subsets of features. the BPSO-NB shown four optimal subsets of features from total of 8 features, and merit value reached to 0.776.

The proposed system used the PID dataset for applying Cross-Validation (CV) and Percentage Split (PS) methods for the classifiers training process and the rest is used for testing or application processes. Data discretization is done to optimize the three classifiers' training process.

In the end of statistical analysis to FS methods based PID dataset, we observed that following features ordered and selected according to optimal and non-optimal features to use. The following Tab.6 presents the sets of features that appeared as a useful feature for optimal performance. The elimination rule is each feature (Fi) which have got zero or near to zero or select features that are much less than others. By evaluation the results of the Tab.2,3,4,5, which showed the selected features, and by applying the exclusion rule, the result became in the optimal and non-optimal features in the Tab.6.

Table 6: FS algorithms and techniques with optimal and non-optimal features

Methods	Optimal Features (Fs)	Non-Optimal Features
Information Gain (IG)	S2,S8,S6,S5,S1	S4,S7,S3
Gain Ratio (GR)	S2,S8,S6,S5,S1	S4,S7,S3
Gini Index (GI)	S2,S8,S6,S5,S1	S4,S7,S3
ANOVA	S2,S8,S6,S5,S7,S1	S4 ,S3
Relief F	S2,S4,S7,S1,S6,S8,S3	S5
FCBF	S2,S8,S6,S7	S5,S1,S4,S3
AUROC	S2,S6,S8,S1,S7,S3	S4,S5
Correlation	S2,S6,S5,S7,S8,S1,S3	S4
Cut-point distance	S2,S5,S1,S6,S7,S8	S4,S3
CFS	S2,S6,S7,S8,	S1,S3,S4,S5
BPSO_NB	S2,S6,S7,S8,	S1,S3,S4,S5

The feature selection results analysis indicates that all the IG, GI, and GR results of the selected features as an optimal and non-optimal feature, where it eliminates three the set of non-optimal features and five the set optimal features. ANOVA test, AUROC curve, and cut-points method indicate to two non-optimal features and six sets of optimal features. While FCBF, BPSO-NB and CFS methods identified a subset of 8 features, which was further cut down by elimination rule methods, resulting in a subset of four optimal and four non-optimal features.

The proposed system effectiveness of the proposed discretize-FSHNB with other classifiers FSNB, FSDT are measured in term of FP Rate, TP Rate, and F-measure; which are calculated based on the Confusion Matrix. The Confusion Matrix is a square matrix where columns correspond to the predicted class, while rows correspond to the real classes. the confusion matrix, which depicts the four possible prediction outcomes, addition to Recall and precision. To obtain the hybrid evaluation by using the following metrics, the proposed method is compared to HNB, NB, and DT:

- Classification accuracy (ACC): it is gained by CV or PS using the selected features ($S_{selected}$) on the test dataset for each classifier. It is calculated by the value of the percentage of correct and incorrect predictions.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (18)$$

- Classification performance: it is gained by Sensitivity, Specificity, Precision, and ROC using the selected features ($S_{selected}$) on the test dataset for each classifier. As shown in the Eq. (19,20,14,17)

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN} \quad (19)$$

$$Specificity\ = \frac{TN}{TN+FP} \quad (20)$$

- Results analysis: Pre-FS classifiers and Post-FS classifiers, the fitness criteria are calculated accuracy and precision, sensitivity, specificity, and ROC.
- Suitability measure: evaluate the accuracies of the three classifiers learned from the optimal subset of features selected by each classifier-based FS techniques and methods, computed by Eq. (16).
- Fitness evaluation: compute and evaluate precision value of PID dataset for each classifier. Whereas, Fitness calculation for selected and obtained features from average of best ranking measures. The fitness function is computed by Eq. (14) .

Basically, the system adopts the PID dataset that is split

in an 80:20 or 70:30 ratio for any modern algorithm, although there are no restrictions on segmentation. The training set is a collection of data that is used to apply various machine learning algorithms to various parameters.

Table 7: Accuracy for each classification-based FS techniques and algorithms

FS Tec and algo.	DT	NB	HNB
Information Gain (IG)	**78.5-CV10F**	**78.8-CV10F**	**82.1-PS79.7**
Gain Ratio (GR)	**78.5-CV10F**	**78.8-CV10F**	**82.1-PS79.7**
Gini Index (GI)	**78.5-CV10F**	**78.8-CV10F**	**82.1-PS79.7**
ANOVA	**78.5-CV10F**	**79-CV10F**	**83.3-PS63.3**
Relief F (RelF)	**75.1-CV10F**	**78.1-CV10F**	**82.5-PS75.4**
FCBF	**79-CV10F**	**79.8-CV10F**	**83.2-PS75.2**
CFS	**82.6-PS75.2**	**82.6-PS75.2**	**83.2-PS75.2**
Correlation	**78.3**	**77.9**	**81.8-PS62.7**
BPSO_NB	**82.6-PS75.2**	**82.6-PS75.2**	**83.2-PS75.2**

*note\ CV10F is Cross Validation with 10Fold dataset, PS is percentage split (Dx Training set, Dn-x Testing set)

For the FS methods applied, the robustness level of HNB model was held constant at around 82%, whereas NB and DT indicate that there is a difference between these methods applied, and the results from all the FS techniques and algorithms from PID dataset were taken into account together.

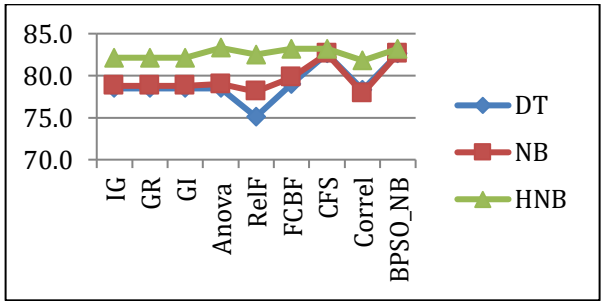

Figure 4: Predicted Accuracy with FS techniques and algorithms

Ideally, we observe that applied FS techniques and algorithms for different subset within training dataset, that show the HNB model is the most stable classifier in terms of model performance, additional it is the best classifier has got classification accuracy than others as shown in the Fig. 4. Whilst, NB and DT is the worst classifiers, when classification models were built using the features selected by the training and testing set, Fig. 4.

Table 8: comparison of the performance measures of PID dataset based on hybrid method for classifiers with FS techniques

Measures	DT	NB	HNB	FSDT	FSNB	FSHNB
Accuracy	**78.26**	**77.86**	**81.8**	**82.6**	**82.6**	**83.3**
Sensitivity	**0.71**	**0.68**	**0.75**	**0.76**	**0.76**	**0.75**
Specificity	**0.81**	**0.83**	**0.845**	**0.85**	**0.85**	**0.87**
F-Score	**0.779**	**0.779**	**0.815**	**0.824**	**0.827**	**0.833**
TP Rate	**0.783**	**0.779**	**0.818**	**0.826**	**0.826**	**0.833**
FP Rate	**0.288**	**0.266**	**0.254**	**0.238**	**0.211**	**0.213**
Precision	**0.778**	**0.779**	**0.815**	**0.824**	**0.828**	**0.833**
ROC	**0.801**	**0.846**	**0.876**	**0.824**	**0.880**	**0.876**

The improved area of the ROC curve proves that the proposed feature selection techniques could enhance the predictive accuracy of the classifier with minimal number of features.

The precision measure maximum as a fitness function is used to quantify the optimality of the features and an efficient fitness function to improve the performance. In the results evaluation of fitness function appears the best results for the FSHNB classifier.

Whilst, when compared precision with the accuracy, accuracy is better for accuracy as a fitness function rather than the precision . otherwise, when too compared with the accuracy, From the results of this analysis it appears that the best measure to use as a fitness function when applying FS techniques and algorithms for data classification are the ROC curve. In some cases the Specificity measure may also be effective. Other significant high-level measures were found , in which the scale of the ROC was shown to appear to be better than specificity by (4:2).

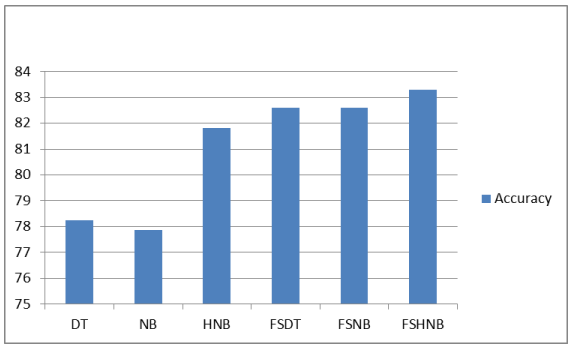

Figure 5: Predictive model of the proposed system based on Hybrid state of the art classification-based FS techniques of PID dataset

Through results analysis of the proposed system, we observed that the feature selection techniques and algorithms to choose the best subsets of PID dataset for using predictive model using state of the art classifiers with high accuracy and performance. Ideal features have been employed using the FS methods, HNB model shows the effective performance than others. Feature selection techniques used on PID dataset have proven that there are subsets of features should deal with their and these techniques have enhanced of HNB model performance and accuracy.

5 Conclusion

The objective of applying BPSO techniques for feature selection with Naive Bayes method for PID dataset classification is a performance test compared with FS methods to get the best technique. The stability, robustness, and fitness aspects of the proposed system are computed and their acceptability. Then, the research also trends to compare performance and the results of the HNB and Naive Bayes and Decision Tree classifications algorithms to find out which algorithm performed well both in quality and accuracy. Our validation results based on the training set indicate that the models-based classification algorithms and discretization-based feature selection techniques perform better. In contrast, the results of the HNB model show that HNB is one of the leading models in terms of performance. Though in accuracy, performance and stability are not necessarily related; however, that HNB model simultaneously attains the highest precision and stability. These approaches work on building full models that can create new features with many adaptive algorithms and classifiers to produce the final results.

Acknowledgment: We acknowledge Universiti Malaysia Pahang (UMP) (PRGS190301) for support related to the proofreading service for this paper: Discretization Method based on Hidden Naive Bayes Algorithm for Diabetes Datasets Classification.

References

[1] J. Brownlee, "Supervised and unsupervised machine learning algorithms," *Mach. Learn. Mastery*, vol. 16, no. 03, 2016.

[2] A. Hayes *et al.*, "Changes in quality of life associated with complications of diabetes: results from the ADVANCE study," *Value Heal.*, vol. 19, no. 1, pp. 36–41, 2016.

[3] J. H. Kamdar, J. J. Praba, and J. J. Georrge, "Artificial Intelligence in Medical Diagnosis: Methods, Algorithms and Applications," in *Machine Learning with Health Care Perspective*, Springer, 2020, pp. 27–37.

[4] V. R. Kebande, R. A. Ikuesan, N. M. Karie, S. Alawadi, K.-K. R. Choo, and A. Al-Dhaqm, "Quantifying the need for supervised machine learning in conducting live forensic analysis of emergent configurations (ECO) in IoT environments," *Forensic Sci. Int. Reports*, vol. 2, p. 100122, 2020.

[5] F. Ali *et al.*, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Inf. Fusion*, vol. 63, pp. 208–222, 2020.

[6] A. Al-Dhaqm *et al.*, "Categorization and Organization of Database Forensic Investigation Processes," *IEEE Access*, vol. 8, pp. 112846–112858, 2020, doi: 10.1109/access.2020.3000747.

[7] S. X. Ding, T. Jeinsch, P. M. Frank, and E. L. Ding, "A unified approach to the optimization of fault detection systems," *Int. J. Adapt. Control Signal Process.*, vol. 14, no. 7, pp. 725–745, 2000.

[8] M. Ngadi, R. Al-Dhaqm, and A. Mohammed, "Detection and prevention of malicious activities on RDBMS relational database management systems," *Int. J. Sci. Eng. Res.*, vol. 3, no. 9, pp. 1–10, 2012.

[9] H. Zhang, P. Lin, J. He, and Y. Chen, "Accurate Strawberry Plant Detection System Based on Low-altitude Remote Sensing and Deep Learning Technologies," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2020, pp. 1–5.

[10] S. J. Park *et al.*, "Development of a real-time stroke detection system for elderly drivers using quad-chamber air cushion and IoT devices," SAE Technical Paper, 2018.

[11] A. Kannan, K. G. Venkatesan, A. Stagkopoulou, S. Li, S. Krishnan, and A. Rahman, "A novel cloud intrusion detection system using feature selection and classification," *Int. J. Intell. Inf. Technol.*, vol. 11, no. 4, pp. 1–15, 2015.

[12] I. U. Onwuegbuzie, S. Abd Razak, I. Fauzi Isnin, T. S. J. Darwish, and A. Al-dhaqm, "Optimized backoff scheme for prioritized data in wireless sensor networks: A class of service approach," *PLoS One*, vol. 15, no. 8, p. e0237154, Aug. 2020, doi: 10.1371/journal.pone.0237154.

[13] L. Jiang, H. Zhang, and Z. Cai, "A novel bayes model: Hidden naive bayes," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 10, pp. 1361–1371, 2009,

doi: 10.1109/TKDE.2008.234.

[14] B.-N. Jiang, X.-Q. Ding, L.-T. Ma, Y. He, T. Wang, and W.-W. Xie, "A hybrid feature selection algorithm: Combination of symmetrical uncertainty and genetic algorithms," in *The second international symposium on optimization and systems biology*, 2008, pp. 152–157.

[15] S. S. Shinde and R. Patil, "Improving spam mail filtering using classification algorithms with discretization filter," *Int. J. Emerg. Technol. Comput. Appl. Sci.*, vol. 10, no. 1, pp. 82–87, 2014.

[16] N. Jatana and K. Sharma, "Bayesian spam classification: Time efficient radix encoded fragmented database approach," in *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, 2014, pp. 939–942.

[17] D. A. Kumar and R. Govindasamy, "Performance and evaluation of classification data mining techniques in diabetes," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1312–1319, 2015.

[18] N. Deepika and S. Poonkuzhali, "Design of hybrid classifier for prediction of diabetes through feature relevance analysis," *Int. J. Innov. Sci. Eng. Technol*, vol. 2, no. 10, pp. 788–793, 2015.

[19] S. Mukherjee and N. Sharma, "Intrusion detection using naive Bayes classifier with feature reduction," *Procedia Technol.*, vol. 4, pp. 119–128, 2012.

[20] P.-P. Wen, S.-P. Shi, H.-D. Xu, L.-N. Wang, and J.-D. Qiu, "Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization," *Bioinformatics*, vol. 32, no. 20, pp. 3107–3115, 2016.

[21] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[22] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*, Elsevier, 1992, pp. 249–256.

[23] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *European conference on machine learning*, 1994, pp. 171–182.

[24] P. Somol and J. Novovičová, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1921–1939, 2010.

[25] M. A. Hall, "Correlation-based feature subset selection for machine learning," *Thesis Submitt. Partial fulfillment Requir. degree Dr. Philos. Univ. Waikato*, 1998.

[26] P. Jafari and F. Azuaje, "An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors," *BMC Med. Inform. Decis. Mak.*, vol. 6, no. 1, pp. 1–8, 2006.

[27] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[28] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation*, 1997, vol. 5, pp. 4104–4108.

[29] N. Kushwaha and M. Pant, "Link based BPSO for feature selection in big data text clustering," *Futur. Gener. Comput. Syst.*, vol. 82, pp. 190–199, 2018.

[30] J. Lin and J. Yu, "Weighted Naive Bayes classification algorithm based on particle swarm optimization," in *2011 IEEE 3rd International Conference on Communication Software and Networks*, 2011, pp. 444–447.

[31] T. GopalaKrishnan and P. Sengottuvelan, "A hybrid PSO with Naïve Bayes classifier for disengagement detection in online learning," *Program*, 2016.

[32] J. Li, L. Ding, and B. Li, "A novel naive bayes classification algorithm based on particle swarm optimization," *Open Autom. Control Syst. J.*, vol. 6, no. 1, 2014.

[33] P. Jongsuebsuk, N. Wattanapongsakorn, and C. Charnsripinyo, "Real-time intrusion detection with fuzzy genetic algorithm," in *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2013, pp. 1–6.

[34] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Appl. Soft Comput.*, vol. 86, p. 105836, 2020.

[35] D. Dernoncourt, B. Hanczar, and J.-D. Zucker, "Analysis of feature selection stability on high dimension and small sample data," *Comput. Stat. Data Anal.*, vol. 71, pp. 681–693, 2014.