Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:113

https://doi.org/10.28919/cmbn/8838

ISSN: 2052-2541

LEAST SQUARES SUPPORT VECTOR MACHINE ENSEMBLE BASED ON SAMPLING FOR CLASSIFICATION OF QUALITY LOCAL CATTLE

BAIN KHUSNUL KHOTIMAH[1,*], EKO SETIAWAN[2], DEVIE ROSA ANAMISA[1], OKTAVIA RAHAYU PUSPITARINI[3], AERI RACHMAD[1]

[1]Department of Informatics Engineering, Faculty of Engineering, University of Trunojoyo Madura, Bangkalan, 69162, Indonesia

[2]Department of Natural Resource Management, Faculty of Agriculture, University of Trunojoyo Madura, Bangkalan, 69162, Indonesia

[3]Department of Animal Husbandry, Islamic University of Malang, Malang, 65144, Indonesia

Abstract: Selection of superior quality local cattle with quality meat with low water and fat content that is very suitable for food processing and supports local wisdom culture. The main problem in selecting superior quality cattle is to choose the right candidate for the parent for breeding with characteristics almost the same as non-local cattle entering Madura. Classification is done to find the best model for selecting superior seeds with unbalanced classes. Using cattle data, this study will apply the LS-SVM ensemble method with combined SMOTE for multi-class imbalanced classification. To overcome high dimensions with unbalanced classes, the gradient Boosting method and sampling technique with SMOTE are applied to balance the number of majority classes into minority classes. The evaluation criteria for classification performance use accuracy values, such as G-means and running time. The experiment used k-fold cross-validation with k=5, with ensemble gradient boosting optimization showing success in improving classification performance. While using kernels, linear kernels produce higher performance and shorter computing time, with the addition of the gradient boosting technique and the best parameters of a σ value of 10 and

*Corresponding author

E-mail address: bain@trunojoyo.ac.id

Received August 14, 2024

C value of 50, and the SMOTE sampling technique produces the highest accuracy of 100%. The addition of gradient boosting has reduced iterations to make faster time on the LS-SVM method, and the correct parameters have produced a Grid Search performance.
Keywords: classification; gradient boosting; sampling technique; superior cattle breeds; LS-SVM.
2020 AMS Subject Classification: 62H30.

1. INTRODUCTION

The general problem in selecting superior quality local Madurese cattle is maintaining excellence to achieve high livestock productivity by the socio-economic conditions of Madura Island [1]. Madura cattle are a type of cattle originating from Indonesia that is a cross between Balinese cattle (Bos sondaicus) and Zebu cattle (Bos indicus) [2]. The identify superior cattle depends on the owner's or community's subjectivity, which still needs a standard pattern. So most people keep cows only because they are hereditary from previously kept cows. The determine quality of cattle requires proper mapping techniques [3]. Cow data based on body condition includes features of body shape characteristics carried out by palpating whether or not there are fat deposits under the skin around the base of the tail, spine and waist, as well as observations based on the color of the fur, horns and hooves, the health of the cow and the rate of weight gain. body based on age [4]. The problem is the difficulty in identifying prospective sires that comply with the optimal standards for Madurese cattle, requiring a model that experts will validate. The selection of superior cattle in Madura has the weakness of not being carried out using advanced technology, so it is subjective because it depends on the expert [5]. The development on a large scale has difficulties and requires continuous regeneration [6]. Cattle data often experiences imbalance because there is a tendency for certain types of cattle to cluster. There are three learning methods and approaches to overcoming class imbalance problems [7][8].

Limitations in determining the body weight of cattle in the field are the lack of livestock weighing facilities so that farmers must make subjective estimates of body weight. Several methods have been developed to predict body weight based on linear body measurements. The method that has been used is the School method which uses chest circumference and the Winter

method using chest circumference and body length as the calculation factors [9]. Calculation using the body condition score (BCS) of livestock is a method that is widely used in the field. This method is simple and easy to use to evaluate nutritional adequacy during the lactation phase. The ideal BCS assessment of livestock depends on the purpose of maintenance. Livestock raised for meat or fattening livestock if the body BCS is larger, the better. Superior seeds Livestock with the aim of prospective parents and enlargement require fat and proportional body conditions. Livestock that are suitable for ideal seeds have a body condition value with livestock that are not too fat and not too thin. Therefore, BCS calculations are very necessary to find out how much nutrition is given so that the condition of the cow is optimal during the next parturition. Body Condition Score is a critical measurement method that affects the milk production of dairy cows, aimed at influencing the performance of production for contests, the more ideal or appropriate the value and reproductive efficiency of prospective sires [10][11].

The first approach is to use level data for Sampling-Based Approach. The second approach is at the algorithm level. The third approach is the ensemble learning method [12]. The sampling approach to imbalanced classes causes the imbalance class level to become smaller and classification can be carried out correctly. The sampling based approach modifies the training data distribution so that both data classes (negative and positive) are represented well in the training data. This resampling technique includes oversampling and under sampling. The most basic methods for dealing with class imbalance problems are Random Over Sampling (ROS) and Random Under Sampling (RUS). The RUS method is done by deleting instances from the majority class while duplicating instances carry out the ROS method from the minority class [13]. Both techniques can handle imbalance class problems. However, both methods have several weaknesses. The ROS method is ineffective in improving the recognition process for minority classes and increasing the classifier formation process time. The RUS method can potentially discard instances from the majority class that are considered important. However, the research states that the RUS method can minimize the negative impact of loss of information while maximizing the positive impact of data cleaning in the undersampling process [14][15][16].

The ensemble learning method is a method that can be applied when a classifier uses more than

one classifier to create a classification model. Meanwhile, the SMOTE algorithm was proposed based on k-nearest neighbors to duplicate minority data to balance the amount of data [14][17]. This research used the SMOTE and Least Square SVM algorithms to provide better results than standard LS-SVM. However, the weakness of the SMOTE algorithm is that it often experiences overfitting. Another research on sampling was carried out by combining sampling between SMOTE and the SVM classifier. As a result, SMOTE sampling provides better accuracy values than without sampling [18].

The SVM method is a machine learning method that is useful and successful in making predictions in both classification and regression cases. The basic principle of SVM is a linear classifier which is then developed for non-linear problems by incorporating the concept of kernel tricks in high-dimensional workspaces. In simple terms, the SVM concept is an attempt to find the best hyperplane that functions as a separator of two classes in the input space [19]. The SVM method was developed based on statistical learning theory and Structural Risk Minimization (SRM), which has shown performance as a method that can overcome the overfitting problem by minimizing the upper limit on generalization error which is a powerful tool for supervised learning cases [20][21]. However, some cases of extreme imbalance combined sampling can perform less well. Another alternative for increasing the accuracy of the imbalance class is to use the ensemble method. The ensemble method in principle combines a group of trained classifiers with the aim of creating an improvised mixed classification model thus making the combined classifier that is formed more accurate than the original classifier in carrying out a classification [19][22]. One very popular ensemble method is boosting, which employs a group of classifiers that are trained iteratively. Gradient Boosting base on Decision Tree (GBDT) in principle forms a strong classifier by combining a group of classifiers, which can expand the margin which can improve generalization capabilities [23]. Gradient Boosting maintains a set of observation weights during observation training and adaptively adjusts (updates) these weights at the end of each boosting iteration. The weights of observations that were incorrectly classified during training will be increased while the weights of observations that are correctly classified will be reduced in value. The SVM method is a machine learning method that is useful and successful in making predictions

in classification and regression cases. The basic principle of SVM is a linear classifier which is then developed for non-linear problems by incorporating the concept of kernel tricks in high-dimensional workspaces. In simple terms, the SVM concept attempts to find the best hyperplane that functions as a separator of two classes in the input space [24][25]. The SVM method was developed based on statistical learning theory and Structural Risk Minimization (SRM), which has shown performance as a method that can overcome the overfitting problem by minimizing the upper limit on generalization error, a powerful tool for supervised learning cases. However, some cases of extreme imbalance combined sampling can perform less well. Another alternative for increasing the accuracy of the imbalance class is to use the ensemble method. The ensemble method in principle combines a group of trained classifiers to create an improvised mixed classification model, thus making the combined classifier that is formed more accurate than the original classifier in carrying out a classification [22][26].

This research uses an ensemble learning method, using Gradient boosting to increase accuracy by shortening the iterations of several of the best parameters. The LS-SVM method was developed to solve problems using the SVM method to overcome lazy learning (high computing) using a Linear equation function, producing a better accuracy level than SVM [27][28]. The condition of imbalanced data is a problem in multi-class classification because the classifier learning machine will tend to predict the large data class (majority) compared to the minority class. As a result, good prediction accuracy is produced for large classes of training data (majority class) while for small training data classes (minority class) poor prediction accuracy will be produced [29][30][33]. Use of sampling techniques, comparison of the use of sampling in multi-class imbalanced data classification and application of Least Square Support Vector Machine (LS-SVM) as preprocessing to overcome imbalanced mixed and imbalanced data. Optimization of RBF γ and C kernel function parameters using Grid Search, adopting previous research [31][32][34].

This approach uses LS-SVM ensemble, by combining sampling techniques to overcome imbalanced data. This research is expected to improve the performance of ensemble classification with linear data, which often experiences overfitting such as cattle data. This paper is organized into several sections as follows. In Section 1, discussing the introduction containing the research

problem and some basic understanding of LS-SVM is reviewed. Section 2, combining LS-SVM with ensemble techniques and individual sampling is explained. Section 3 and section 4, the results and discussion of some simulation results are illustrated. Finally, Section 5 draws some conclusions.

2. RESEARCH METHODOLOGY

2.1 Dataset Description

Primary data is collected directly from local contests and beef cattle breeders for economic security. The data consists of x data features, the main criteria for the cow's BCS. Figure 1 measured body shape by palpation for the presence or absence of fat deposits under the skin around the base of the tail, spine, and waist, as well as observations based on the color of the fur, horns, and hooves, the health of the cow, and the rate of body weight gain based on age.

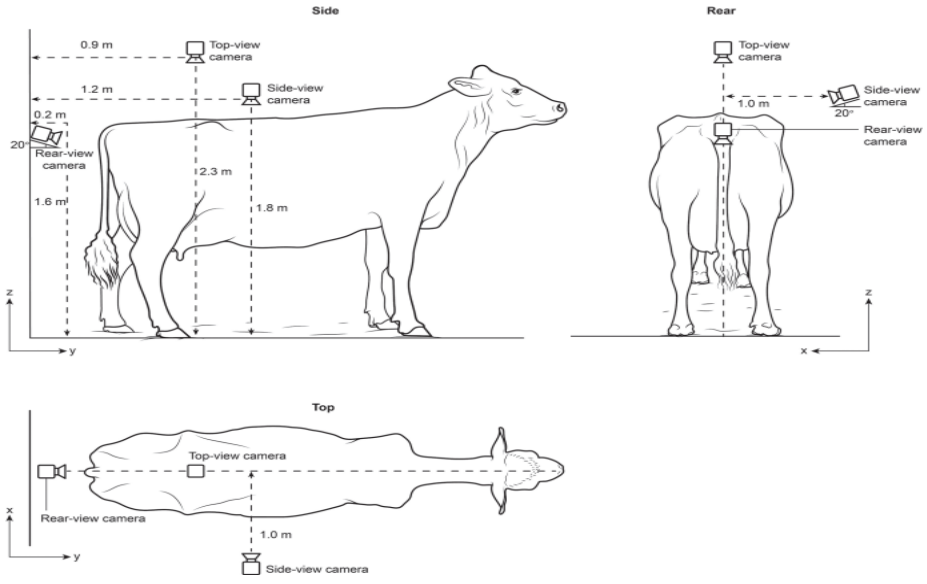

Figure 1. Measurement of Body Condition Score (BCS) of Cattle [14]

Table 1. The results of measuring cow body condition features

No.	Variable	Pattern		
		I (Sonok cattle)	II (Karapan cattle)	III (beef cattle)
1.	Breast width (cm)	143.31±11.49	140.38±8.49	138.20±6.50
2.	Hip height (cm)	118.05±7.26	117.25±9.50	114.75±13.50
3.	Abdominal Circumference (cm)	53.60±1.41	60.19±3.19	50.87±2.59
4.	Head Length (cm)	41±3.69	40±4.55	38±3.41
5.	Head width (cm)	19±3.75	18±2.27	17±2.08
6.	length in chest (cm)	16±3.52	15±5.30	14±2.52
7.	Hight (cm)	120.19±9.74	125±17.05	118±11.50
8.	body long (cm)	134.13±7	123±13.07	120±13
9.	shoulder high (cm)	135±61	128±4	121±5
10.	chest size (cm)	166±5	161±164	154-167
11.	Gumba height (cm)	121 ± 3,43	123 ± 3,45	126 ± 4,10
12.	weight (kg)	300. 34 ±100.50	275±100.50	200±300.50

**Different letters in the same row indicate highly significant differences (P<0.01); (n=365).

Cow data patterns are divided into 3, namely as Cattle Class (y): 1=" Karapan Cattle," n1=70 (24%); 2=" Sonok Cattle," n1=85 (26%); 3=" beef Cattle," n1=210 (56%). Data on local Madurese cattle was taken at ages 24-60 (months), consisting of 12 features with three classes, considered productive cattle ready for consumption. Body condition assessment (BCS) of cattle allows the evaluation of different cattle fat reserves. Measurements are made in productive cattle when evaluated at key production time points. Body size indications, constantly changing according to periodic standardization, have shown that BCS is accurate and helpful on a herd basis. The test carried out is a normality test carried out to determine whether the data has a standard or abnormal distribution to produce a hypothesis [20], with chi testing carried out by calculating X^2_{hitung}:

$$\chi^2{}_{hitung} = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right) \quad (1)$$

The Pearson correlation test is used to statistically test whether the correlation between these variables is significantly different from zero or not. The null hypothesis ($H0$) states that there is no correlation between variables, while the alternative hypothesis ($H1$) states that there is a correlation between variables. Meanwhile, the alternative hypothesis ($H1$) states that there is a correlation between these variables, for potential overfitting in our model. Correlation testing between Manual BCS and observed behavioral features based on observations and historical history of cow track records as measured by correlation determination according to equation [20]:

$$r^2 = \frac{[\, n \sum X_i Y_i - (\sum X_i)(\sum Y_i)\,]^2}{\sqrt{\{n(\sum X_i^2) - (\sum X_i)^2\}\{n(\sum Y_i^2) - (\sum Y_i)^2\}}} \quad (2)$$

Description of correlation test data to determine the ranking of feature influence and testing the level of significance for each indicator variable can be seen in Table 2.

Table 2. The results of testing the assumption of the proportional data variable cows

No.	Variable	r	P-Value	Description
1.	Breast width (cm)	0,3123	0,5392	Significant
2.	Hip height (cm)	-0,0290	0,5389	Significant
3.	Abdominal Circumference (cm)	-0,0692	0,03350	No Significant
4.	Head Length (cm)	0,0283	0,9671	Significant
5.	Head width (cm)	0,7164	0,4854	Significant
6.	length in chest (cm)	-0,1564	0,0249	No Significant
7.	Hight (cm)	0,4276	0,3209	Significant
8.	body long (cm)	-0,5212	0,3479	Significant
9.	shoulder high (cm)	0,0382	0,08471	No Significant
10.	chest size (cm)	0,0011	0,6955	Significant
11.	Gumba height (cm)	0,3209	0,7885	Significant
12.	weight (kg)	0,6129	0,8574	Significant

The following are the results of the analysis that performs a partial test using Spearman rho correlation, which is a statistical method used to test the assumption of a relationship between variables if the data is ordinal (ranking). Based on the simultaneous test, the p-value is obtained less than $\alpha = 0.05$, then reject *H0*, which means that there is one variable that significantly affects the quality of superior Madura cattle. The r value of the relationship between the two variables, namely length in chest and shoulder high, namely the variables, does not meet (violate) the assumption, modeling ten significant variables as significant variables.

2.2 Syntetics Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method, namely a sampling technique to increase the amount of data in the minority class by randomly replicating the amount of data in the minority class so that the amount is the same as the data in the majority class [13][15]. The SMOTE algorithm using this approach generates "synthetic" data, namely new data replicated from minor data. The SMOTE algorithm looks for nearest neighbors and groups data based on nearest neighbors. The closest neighbors were selected based on the Euclidean Distance between the two data sets [16]. Suppose given data with p variables as $x^T = [x_1, x_2, \dots., x_p]$ and $z^T = [z_1, z_2, \dots., z_p]$, then the Euclidean Distance $d(x, z)$ is generally in equation (2.1).

$$d(x, y) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_p - z_p)^2} \tag{2.1}$$

Synthetic data generation is done using equation (2.2):

$$x_{syn} = x_i + (x_{knn} - x_i)x\gamma \tag{2.2}$$

This method will produce new synthetic data that will be used as training data for the classification process, using the K-Nearest Neighbor (KNN) method.

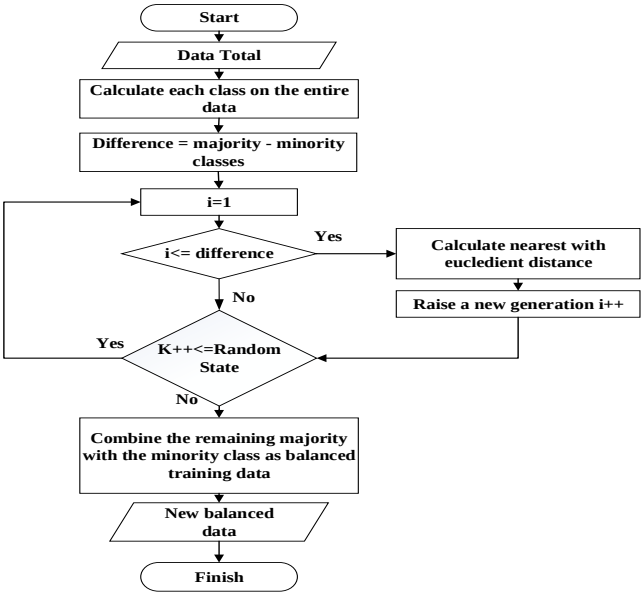

Figure 2. The process of balancing data using SMOTE

2.3 Ensemble Learning

The approach assigns different weights to each method using the least squares method, which optimizes the contribution of each particular individual estimate [17]. Several ensemble learning methods are widely used, such as Boosting, Bagging, and random forest [18]. Boosting is an approach to machine learning to increase accurate predictions by combining many weak and inaccurate rules. Adaptive boosting (AdaBoost) is one of several variants of the boosting algorithm, which is generally combined with a classifier to improve classification performance. Boosting technology is a machine learning technology with tree optimization, where the objective function is simplified by combining prediction and regularization terms while maintaining the Fastest possible processing speed. Gradient Boosting using Decision Tree (GBDT) uses scores containing information about the specific threshold used for classification. GB uses decision trees to learn a function from the input space Xs to the gradient space G [21]. Suppose we have a training set with n i.d. instances $\{x1, \ldots, xn\}$, where each xi is a vector of dimension s in the Xs space.

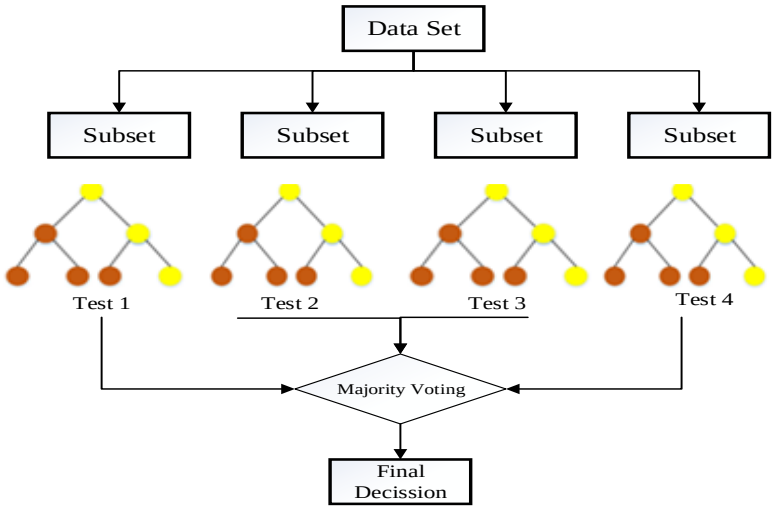

Figure 3. Gradient Boosting Decision Tree (GBDT) Diagram

Figure 3 analytics utilizes voting GBDT techniques according to the equation to reduce mathematical speed and prevent overfitting by the model. Establish a basic model to predict a data set by taking the average of each target column according to the equation (2.3). This value anticipates a representative overall value by taking the highest vote in a particular iteration to prevent overfitting [27].

$$F_0(x) = arg_\gamma min \sum_{i=1}^{n} L(y_i, \gamma) \tag{2.3}$$

If the output value is from each leaf in the decision tree, to find the output of all the leaves because one leaf can produce more than one residue. The result is calculated by taking the average of all values in one leaf.

$$\gamma_m = arg_\gamma min \sum_{i=1}^{n} L\left(y_{i,} F_{m-1}(X_i) + \gamma h_m(X_i)\right) \tag{2.4}$$

Where, L present Loss Function, and γ is Gamma as Predicted Value argmin= predicted value or gamma to be found of which the loss function is minimum. Since, update for next model base iteration.

$$Fm(X) = F_{m-1}(X) + v_m h_m(X) \tag{2.5}$$

where M presents no. of decision trees made, Fm-1(x)= forecasts of the base model. Utilize the test set's predictions as features for the meta-model, a new model. Utilizing the meta model, make final predictions for the test set.

2.4 Grid Search (GS)

Grid Search (GS) trains a machine-learning algorithm for all combinations of hyperparameters. The performance measurement process uses a cross-validation technique on the training set to generate models. Grid Search calculates each model's score, evaluates them, and then selects the model that gives the best results.GS is to exploit the search in the number of evaluations that increases exponentially as the frequency of hyperparameters increases. Assuming that *k* parameters have n grid-separated values, then its computational complexity increases exponentially at the rate of $O(n_k)$ [26]. Thus, GS can only be an efficient HPO approach when the hyperparameter configuration space is limited [30].

2.5 Building an Ensemble LS-SVM Model with Sampling Techniques

This stage is to build Optimization on LS-SVM applied to the imbalance data obtained from validated data collection. Optimization is done using the Gradient Boosting method, abbreviated as Gboost LS-SVM, based on sampling. The stages of model development are described as follows:

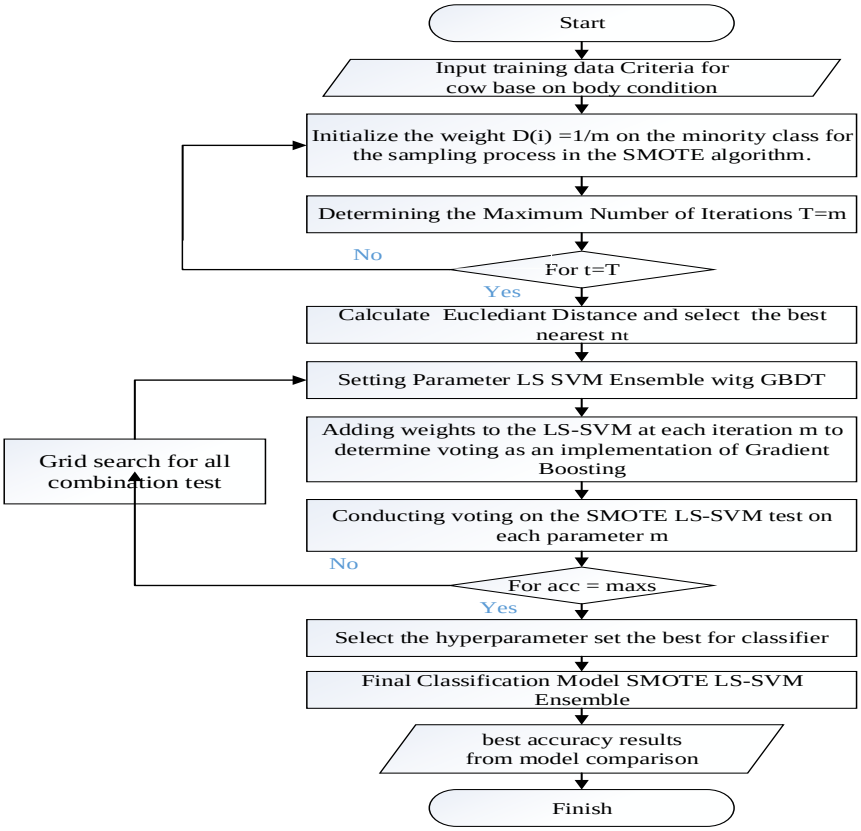

Figure 4. LS-SVM Ensemble Flow Diagram based on sampling technique

Figure 4. explains the hybrid steps of the proposed combined method developing an LS-SVM model with the addition of Gradient boosting (SMOTEGboost LS-SVM) the following stages:

1. Set data balance with the SMOTE algorithm

   SMOTE to duplicate the minority class into the same class as the majority. The SMOTE algorithm adds weight to the training data with a value of $D_1(i)$ =1/m, where m is the number of training data observations.

2. Determine the LS-SVM Model

   Determine the range of parameters C and RBF kernel parameters γ that will be optimized with grid search. Parameter C is set between 0.1, 1, 10, 100 and parameter γ between 0.01, 0.1, 1, 10. Linear and nonlinear transformations, Φ are needed to map data from the original feature space to a new, higher dimensional space.

   The SVM was developed into a Least Square-Support Vector Machine method to reduce computation, because the use of variables LS-SVM only uses two variables (γ and $\sigma 2$) less than SVM which uses three variables (C, $\sigma 2$, and Ɛ) [23], according to the following equation (2.6) and (2.7):

$$\frac{1}{2}\omega^2 + \frac{1}{2}\gamma \sum_{i=1}^{n} e_i^2 \tag{2.6}$$

$$y_i(\omega^T x_i + b) + e_i = 1, i = 1, \dots, n \tag{2.7}$$

   The existence of these changes is done by changing the Lagrange Function with the following equation (2.8):

$$(L, \omega, b, e, \alpha) = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{n} \alpha_i [y_i(\omega^T x_i + b) + e_i - 1] \tag{2.8}$$

   γ = Gamma (regulatory factors), x_i = Input Data, ω = Weight value, b = The bias value that needs to be changed to unconstrained optimization, and y_i = Output. The application of the gradient boosting algorithm works sequentially by combining learning results from decisions from various learning machines combined, resulting in the best learning. The flow of the Gradient Boosting algorithm is as follows:

1.Initialize $f_0(x)$ tosingle terminalnode tree

2. For $m = 1$ to M :

a) Compute pseudo residuals r_{im} based on loss function

b) Fit a regression tree to $r_{im} \Rightarrow R_{jm}$, $j = 1,2,\ldots J_m$

c) Find the optimal value of coefficient within different region R_{lm}

$$\gamma_{jm} = \arg\min_{\gamma_{jm}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

$$d)\ f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

End For

Output : $\hat{f} = f_M$

2.6 Evaluation Measures

Actual data and predicted data from the classification model are presented using Cross Tabulation (Confusion Matrix), which contains information about the actual data class represented in the rows of the matrix and the predicted data class in the columns [23].

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.9}$$

Sensitivity and Specificity can be tested for an optimal and more specific classification. Sensitivity is the true positive rate or performance measure to measure the positive (minor) class, while Specificity is the true negative rate or performance measure to measure the negative (major) class [24]. The Sensitivity and Specificity formulas are as follows.

$$Sensitivity = \frac{TP}{(TP+FN)} x100\% \tag{2.10}$$

$$Specifity = \frac{TN}{(TN+FP)} x100\% \tag{2.11}$$

In addition, the performance of the classification model can be evaluated using G-mean and F-measure. If all positive classes cannot be predicted, then the G-mean will be zero, so a classification algorithm is expected to achieve a high G-mean value with formula (13).

$$G - Mean = \sum Sensitivity\ x\ Specifity \tag{2.12}$$

3. MAIN RESULTS

3.1 Replication of inbalanced data with SMOTE

The SMOTE method is an oversampling method used to increase the number of minority classes by randomly replicating data according to the desired percentage so that the number approaches the number of major data. The application of the oversampling method to imbalanced data causes the level of imbalanced data to be smaller and classification can be done correctly. The results of handling the SMOTE method on each imbalanced data used in this study are shown in Table 3.

Tabel 3. Description of Data Distribution Before and After SMOTE

Previous Data		Replication	Data after
Kelas Mayor	Kelas Minor		
Kelas 1=300;	Kelas 2=66;	3 tahap	(66*)(48%**)3**
	Kelas 3=80;	2 tahap	(80*)(34%**)2**

This study uses local cattle data from the Madura area, famous for its cattle for shows and races. Each piece of data in class 2 and class 3 will be replicated so that the amount will increase and the data in the significant class will be balanced. The maximum number of nearest neighbors, meaning replication stage 1, is carried out where class 2 is 11 times on the data, so the amount of new data class 2 has 66 data while the amount of significant data is 300. This condition still has a massive difference with the amount of significant data, so 3-stage replication is carried out for cases like this. Since class 3, totaling 80 data, will carry out 2-stage replication, with eight replications with a nearest of 10 to class 3, a significant class with 300 members, no replication is carried out.

3.2 Classification Using LS-SVM

LS-SVM method with the use of RBF kernel (σ) and (C) using values in the specified range. Determining the range will determine the accuracy. The RBF kernel parameters (σ) that were tried were (range 1-20) and the parameters (C) that were tried were (range 1-100). The classification

results using LS-SVM used 5 Folds.

Table 4. Test results using Multiclass LS-SVM classification

C	σ	Accuracy(%)	Sensitivity(%)	Specificity(%)	N-Gram(%)
1	1	98.34	88.20	93.30	98.72
	10	97.47	75.34	90.21	94.34
	20	96.20	74.04	88.30	92.36
20	1	96.12	96.00	86.41	93.04
	10	99.34	77.34	78.23	94.34
	20	95.10	92.00	78.20	98.34
50	1	96.09	93.26	68.04	97.24
	10	100.00	85.00	83.12	97.36
	20	96.30	98.34	83.34	93.82
100	1	100	93.31	75.30	96.30
	10	95.90	84.30	94.27	91.04
	20	96.30	98.32	97.20	98.28

Table 4 shows that in the cattle data, the highest average percentage of classification accuracy training is 100% when using a value σ of 10 and a value of C of 50. The highest average percentage value of classification sensitivity is 85.00%, specificity is 83.12%, and N-Gram value is 97.36%. A high accuracy value does not necessarily mean that N-Gram has a maximum value because it is very susceptible to parameter selection.

3.3 LS-SVM Multiclass Classification with the addition of Gradient Boosting Decision Tree (GBDT)

The usage learning iteration has a gradient boosting using voting, denoted as $\{g_1, \ldots, gn\}$. The decision tree model divides each node into the most informative features (with the most significant information gain). This section will explain the results of applying grid search and gradient boosting to the cattle dataset. Gradient boosting carries the working principle of gradient descent combined with the boosting technique. The function of the learning rate is to change the gradient value using the scalar function [31]. The parameter n_estimator is the total number of trees formed

by each subset of data [34]. The max_depth parameter is the depth of the tree formed and has a different value, so determining max_depth will regulate the maximum number of depths formed by the tree. In addition, max_depth aims to prevent overfitting [35]. Machine learning algorithms have hyperparameters, and predictive performance is greatly influenced by determining the number of learning rates, min sample split, max depth, max feature and subsample, and many other parameters related to the model fitting process. The result of LS-SVM gradient boosting is shown in Table 5.

Table 5. Grid Search Experiment results on GBDT for LS-SVM method

Parameters	Tuning GS	Default value	Accuracy (%)	Sensitivity (%)	Specificity (%)	N-Gram(%)
Learning rate	0.025, 0.05, 0.1, 0.2, 0.3	0.1				
N_estimator	50, 100, 250, 300, 400, 500					
Min sample split	2, 5, 10, 20, 30	5	100	97.89	95.24	81.56
Max depth	2, 3, 5, 7, 9, unlimited	3				
Max feature	Log2, sqrt, 0.25, 0.5, 1.0	1.0				
Sub sample	0.15, 0.5, 0.75, 1.0	0.75				

Table 5 shows that optimization with hyperparameter tuning using Grid Search on machine learning models makes the model selection process more manageable. Grid Search uses cross-validation k=5 for each model parameter without manually validating them. When combined with good understanding and intuition, it will provide accurate and optimal predictions. The experimental results show that the GBDT model obtained the best value of 100%.

Further research shows that grid search could be more robust in the tuning process for differences in the results of testing classification methods, which takes time when hyperparameters

are added because the number of parameter combinations increases exponentially. The comparison of the methods of each of the best parameters is shown in Table 6.

Table 6. Differences in the results of testing classification methods

Method	C	σ	Accuracy(%)	Sensitivity(%)	Specificity	N-Gram
LS-SVM	1	1	98.34	98.20	72.34	92.30
LS-SVM GBDT	5	10	97.47	77.68	65.02	97.04
LS-SVM SMOTE	10	20	96.20	68.41	98.34	93.34
LS-SVM SMOTE GBDT	50	10	100	98.50	98.26	98.34

Table 6 shows the results of the different methods with better accuracy results. The usage parameter of C =50, σ=20, with optimal learning rate obtained at a value of 0.1 before, even though there are several smaller values, this does not necessarily apply to different datasets. The max_depth means that the number of trees produced is also not directly proportional to the accuracy value for LS-SVM SMOTE GBDT created until 100%, with the best n_estimator = 300. The testing process has the best value in combining ensemble and hyperparameters on balanced data to achieve optimum accuracy.

4. DISCUSSIONs

Testing for time complexity on LS-SVM and some combinations is an interpretable model with high transparency, while Gradient Boosting Trees are more complex and challenging to interpret. The performance of Gradient Boosting Trees requires considering the trade-off between model performance and interpretability. So, to determine the method's performance, the calculation of the execution time is carried out at the highest accuracy position on the resulting model, as shown in Table 7.

Table 7. Testing results of several methods based on the best parameters and execution time.

Method	C	σ	Time Execute (s)
LS-SVM	1	1	58
LS-SVM GBDT	1	10	63
LS-SVM SMOTE	20	20	42
LS-SVM SMOTE GBDT	20	1	65

Furthermore, the performance of cattle data classification based on iteration termination shows that the accuracy, sensitivity, specificity, and g-mean values tend to reach maximum values at different iterations up to 32. Increasing the number of iterations will produce the highest g-mean, accuracy, sensitivity, and specificity performance. When the number of iterations increases, the computation time also increases. The performance curve based on accuracy for each number of iterations used is shown in Figure 5.

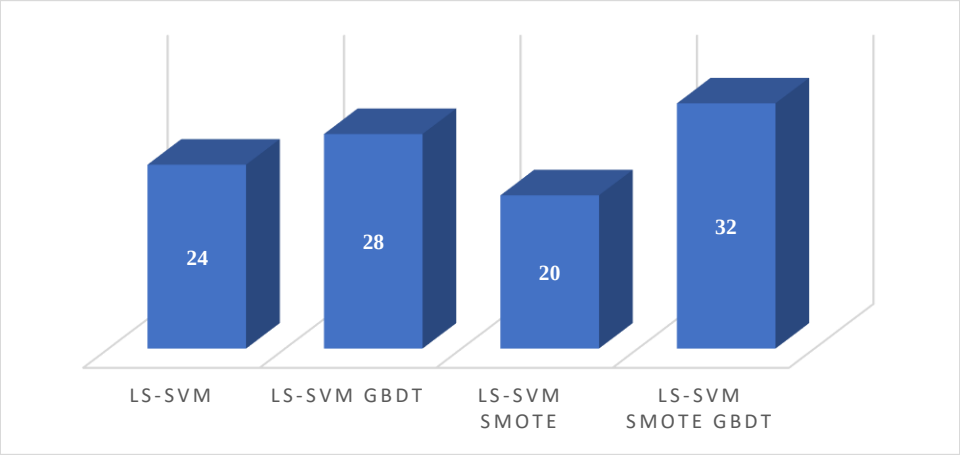

Figure 5. Results of iteration termination on several methods

Figure 5 shows the results of testing several methods, namely LS-SVM GBDT, LS-SVM SMOTE, and LS-SVM SMOTE GBDT, by comparing the performance of all optimal models to determine the final iteration gain. GBDT successfully overcomes the classification errors made by LS-SVM at each iteration to boost and increase classification accuracy in balanced classes. Finally, the addition of SMOTE and GBDT increases classification accuracy, although it produces the highest time at each boosting iteration.

5. CONCLUSIONS AND SUGGESTION

This study utilizes the Sampling technique with SMOTE then continued with the unbalanced data preprocessing process. They are making improvements from minor classes to significant classes by calculating the Euclidean distance by replicating several stages of minor data. The LS-SVM Sampling method with Grid Search hyperparameter optimization is the best method for unbalanced classification cases in predicting superior cattle. Classification using 5-Fold produces performance. The LS-SVM and LS-SVM SMOTE methods have made the best parameters for the C and σ values for LS-SVM. When using the radial kernel, the accuracy obtained in classifying cattle data is 100%, the most considerable G-Mean value. The sensitivity value obtained shows that the separator function obtained can detect 98.50% of observations from the regular class correctly. Observations are correctly classified using the radial kernel. Based on the review above, parameter tuning in LS-SVM can increase the accuracy value but does not always increase Precision, Recall, and N-Gram. Determining the search space value will affect the parameter tuning results, resulting in high accuracy.

The next research suggestion is to improve LS-SVM, in addition to using hyperparameters, ensembles, and sampling. The method requires kernel improvements such as Affinity Propagation (AP), KNN clustering, etc. The linear kernel replacement method is likely more effective and can improve accuracy.

FUNDING

Thank you to all parties for Trunojoyo University, Madura, and those who funded our research in the 2024 budget year. This program is the result of collaboration between Trunojoyo University, Madura, and related parties, namely the DKPP in the Madura region, especially in the field of animal husbandry and the Islamic University of Malang, as a form of cooperation in assisting facilities and infrastructure so that this research can be carried out correctly.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] S.B. Siswijono, P.S. Winarto, R. Prafitri, Strategy for improving production performance and preservation of madura cattle, IOP Conf. Ser.: Earth Environ. Sci. 478 (2020) 012072. https://doi.org/10.1088/1755-1315/478/1/012072.

[2] B.K. Khotimah, F. Agustina, O.R. Puspitarini, et al. Random Search Hyperparameter Optimization for BPNN to Forecasting Cattle Population, E3S Web Conf. 499 (2024), 01017. https://doi.org/10.1051/e3sconf/202449901017.

[3] B.K. Khotimah, F. Agustina, O.R. Puspitarini, et al. Hyperparameters and centroid improvements in the K-medoids method for grouping processed beef SMEs, Commun. Math. Biol. Neurosci. 2024 (2024), 13. https://doi.org/10.28919/cmbn/8369.

[4] P. Balasso, G. Marchesini, N. Ughelini, et al. Machine learning to detect posture and behavior in dairy cows: Information from an accelerometer on the animal's left flank, Animals 11 (2021), 2972. https://doi.org/10.3390/ani11102972.

[5] A.M. Siregar, Y.A. Purwanto, S.H. Wijaya, et al. Comparison of dairy cow on morphological image segmentation model with support vector machine classification, J. RESTI (Rekayasa Sist. Teknol. Inf.) 6 (2022), 670–676. https://doi.org/10.29207/resti.v6i4.4156.

[6] X. Song, E.A.M. Bokkers, S. van Mourik, et al. Automated body condition scoring of dairy cows using 3-dimensional feature extraction from multiple body regions, J. Dairy Sci. 102 (2019), 4294–4308. https://doi.org/10.3168/jds.2018-15238.

[7] H. Djellali, N. Ghoualmi-Zine, S. Guessoum, Hybrid adapted fast correlation FCBF-support vector machine recursive feature elimination for feature selection, Intell. Decis. Technol. 14 (2020), 269–279. https://doi.org/10.3233/idt-190014.

[8] L. Yu, 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), pp. 856-863, (2003).

[9] A. Paul, C. Bhakat, S. Mondal, et al. Body condition score is not a predictor of back fat in primiparous crossbred cattle, Int. J. Basic Appl. Biol. 5 (2018), 45-47.

[10] A. Paul, C. Bhakat, S. Mondal, et al. An observational study investigating uniformity of manual body condition scoring in dairy cows, Indian J. Dairy Sci. 73 (2020), 77–80. https://doi.org/10.33785/ijds.2020.v73i01.013.

[11] X. Song, E.A.M. Bokkers, S. van Mourik, et al. Automated body condition scoring of dairy cows using 3-dimensional feature extraction from multiple body regions, J. Dairy Sci. 102 (2019), 4294–4308. https://doi.org/10.3168/jds.2018-15238.

[12] J. Zhang, X. Cui, J. Li, et al. Imbalanced classification of mental workload using a cost-sensitive majority weighted minority oversampling strategy, Cognit. Technol. Work 19 (2017), 633–653.

https://doi.org/10.1007/s10111-017-0447-x.

[13] A. Hanskunatai, A new hybrid sampling approach for classification of imbalanced datasets, in: 2018 3rd International Conference on Computer and Communication Systems (ICCCS), IEEE, Nagoya, 2018: pp. 67–71. https://doi.org/10.1109/CCOMS.2018.8463228.

[14] H. Dharmawan, B. Sartono, A. Kurnia, et al. A study of machine learning algorithms to measure the feature importance in class-imbalance data of food insecurity cases in Indonesia, Commun. Math. Biol. Neurosci. 2022 (2022), 101. https://doi.org/10.28919/cmbn/7636.

[15] H. Hairani, D. Priyanto, A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data, Int. J. Adv. Computer Sci. Appl. 14 (2023), 585-590. https://doi.org/10.14569/ijacsa.2023.0140864.

[16] T. Phillips, W. Abdulla, Developing a new ensemble approach with multi-class SVMs for Manuka honey quality classification, Appl. Soft Comput. 111 (2021), 107710. https://doi.org/10.1016/j.asoc.2021.107710.

[17] W. Kim, J. Park, H.J. Kim, Target localization using ensemble support vector regression in wireless sensor networks, in: 2010 IEEE Wireless Communication and Networking Conference, IEEE, Sydney, Australia, 2010: pp. 1–5. https://doi.org/10.1109/WCNC.2010.5506589.

[18] D.R. Anamisa, A. Jauhari, F.A. Mufarroha, Performance test of Naive Bayes and SVM methods on classification of malnutrition status in children, Commun. Math. Biol. Neurosci. 2024 (2024), 25. https://doi.org/10.28919/cmbn/8429.

[19] H. Núñez, L. Gonzalez-Abril, C. Angulo, Improving SVM classification on imbalanced datasets by introducing a new bias, J. Classif. 34 (2017), 427–443. https://doi.org/10.1007/s00357-017-9242-x.

[20] R. Rofik, R.A. Hakim, J. Unjung, et al. Optimization of SVM and gradient boosting models using GridSearchCV in detecting fake job postings, Matrik: J. Manaj. Tek. Inf. Rekayasa Komput. 23 (2024), 419–430. https://doi.org/10.30812/matrik.v23i2.3566.

[21] H. Ibrahim, S.A. Anwar, M.I. Ahmad, Classification of imbalanced data using support vector machine and rough set theory: A review, J. Phys.: Conf. Ser. 1878 (2021), 012054. https://doi.org/10.1088/1742-6596/1878/1/012054.

[22] W. Zhang, C. Li, B. Zhong, LSSVM parameters optimizing and non-linear system prediction based on cross validation, in: 2009 Fifth International Conference on Natural Computation, IEEE, Tianjian, China, 2009: pp. 531–535. https://doi.org/10.1109/ICNC.2009.26.

[23] A. Lawi, F. Aziz, Classification of credit card default clients using LS-SVM ensemble, in: 2018 Third International Conference on Informatics and Computing (ICIC), IEEE, Palembang, Indonesia, 2018: pp. 1–4. https://doi.org/10.1109/IAC.2018.8780427.

[24] L. Zhou, K.K. Lai, L. Yu, Least squares support vector machines ensemble models for credit scoring, Expert Syst. Appl. 37 (2010), 127–133. https://doi.org/10.1016/j.eswa.2009.05.024.

[25] Z. Wang, Z. Zhang, J. Mao, Adaptive tracking control based on online LS-SVM identifier for unknown nonlinear system, in: 2012 IEEE International Conference on Information Science and Technology, IEEE, Wuhan, China, 2012: pp. 112–117. https://doi.org/10.1109/ICIST.2012.6221618.

[26] N.R. Goluguri, S. Devi K, P. CH, Infectious diseases of Rice plants classified using a deep learning-powered least squares support vector machine model, Indian J. Computer Sci. Eng. 13 (2022), 1640–1659. https://doi.org/10.21817/indjcse/2022/v13i5/221305186.

[27] M. Farasat, M. Seyedian, K. Daab, Evaporation modeling of free surface water using SVM and LSSVM models, J. Irrigation Water Eng. 11 (2021), 272-288.

[28] D.M. Belete, M.D. Huchaiah, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results, Int. J. Computers Appl. 44 (2021), 875–886. https://doi.org/10.1080/1206212x.2021.1974663.

[29] L. Nanni, A. Lumini, An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring, Expert Syst. Appl. 36 (2009), 3028–3033. https://doi.org/10.1016/j.eswa.2008.01.018.

[30] L. Zhou, K.K. Lai, L. Yu, Least squares support vector machines ensemble models for credit scoring, Expert Syst. Appl. 37 (2010), 127–133. https://doi.org/10.1016/j.eswa.2009.05.024.

[31] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing 415 (2020), 295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

[32] H.A. Fayed, A.F. Atiya, Speed up grid-search for parameter selection of support vector machines, Appl. Soft Comput. 80 (2019), 202–210. https://doi.org/10.1016/j.asoc.2019.03.037.

[33] M. Ogunsanya, J. Isichei, S. Desai, Grid search hyperparameter tuning in additive manufacturing processes, Manuf. Lett. 35 (2023), 1031–1042. https://doi.org/10.1016/j.mfglet.2023.08.056.

[34] A.S. Mohammed, Ş.E. Amrahov, F.V. Çelebi, Interpolated binary search: An efficient hybrid search algorithm on ordered datasets, Eng. Sci. Technol. Int. J. 24 (2021) 1072–1079. https://doi.org/10.1016/j.jestch.2021.02.009.