Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

An Efficient Categorization of Diabetes Imbalanced Data Using SMOTE-ENN With Fine-Tuned LS-SVM Algorithm

[1]Nwayyin Najat Mohammed
Computer Science Department
University of Sulaimani
Sulaimani, Iraq,
naween.mohammed@univsul.edu.iq

[2]Mariwan Hama Saeed
College of basic education
University of Halabja
Halabja, Iraq,
mariwan.ahmedh@gmail.com

ARTICLEINFO
Article History
Received: 30/03/2025
Accepted: 25/04/2025
Published: 28/06/2025

ABSTRACT

Diabetes has been recognized as a major cause of death. Diabetes is a chronic disease. In recent years, the impact of diabetes has increased dramatically, and it has become a global threat. Machine learning is a part of computational algorithms designed to imitate human intelligence by learning from the surrounding environment. Type 2 diabetes is indicated by deviation high blood glucose levels attributable to insulin resistance and reduced pancreatic insulin production. In this study, two diabetes datasets are used, the Pima Indians diabetes and Iraqi Society Diabetes(ISD)datasets. They are collection of data on diabetes which characterized by an imbalanced distribution and the presence of outliers. The diabetes data sets are preprocessed. Many methods, including data resampling have been proposed to address the data sets imbalance issue. We utilized the resampling Synthetic Minority Oversampling and Edited Nearest Neighbors(SMOTE-ENN) technique to address the imbalance diabetes datasets issue and imputation. The classification of imbalanced datasets is a crucial field in machine learning. The machine learning approach that is used in this study is the Least Square Support Vector Machine (LS-SVM) to categorize the diabetes patients. Machine Learning (ML) algorithms are constructed by a set of hyperparameters. Thus, hyperparameters values should be carefully chosen. We used grid search algorithm to optimize LS-SVM algorithm hyperparameters. The classification results were improved. In addition, we could enhance the performance of the fine-tuned LS-SVM with the used resampling technique, SMOTE-ENN, that processes diabetes datasets. The performance metrics that evaluate the proposed algorithm SMOTE-ENN and fine-tuned LS-SVM are accuracy, recall and precision. The metrics measurements obtained were much better and higher when the proposed algorithm was used to

Keywords: *Diabetes Mellitus; Imbalanced dataset; Preprocessing; Resampling; SMOTE-ENN; Least square support vector machine; Optimization.*

1. INTRODUCTION

The Electronic Health Records EHRs data is used for clinical investigation. It contains a huge amount of information about an individual's health status, which are demographics, diagnoses, laboratory test results, high-frequency physiological waveform signals, and others. Data analysis approaches can be used to extract useful models when a sufficient amount of EHR data is collected. The Electronic Health Records EHRs data can provide decision support technologies to assist clinicians in providing better care. This process can promote the development of decision support systems by validating a logical framework. The attributes of Electronic Health Records EHRs data may not be optimal for data analysis [1] [2]. For example, we used the Pima Indian diabetes dataset which is imbalanced dataset. Raw data usually contains many shortages, such as inconsistencies, missing values, noise and redundant information. The performance of learning models will be affected if they are developed with poor quality data. Thus, the preprocessing techniques significantly impact on the quality and reliability of learning models [3]. The Pima Indian diabetes dataset contains entries with a value of zero. These entries are managed using imputation as the first preprocessing approach. Simple imputation techniques include the mean, median and mode. The median imputation is used in this study [4]. The imbalanced datasets occur in many real-world domains.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

The imbalanced class distribution of a dataset is difficult for most learning algorithms to address, as it is assumed that there is a balanced class distribution [5]. The Resampling techniques, over- and under-sampling, have received notable attention for their ability to address imbalanced datasets. Therefore, the second preprocess technique used in this work is the hybrid Synthetic Minority Oversampling and Edited Nearest Neighbors SMOTE-ENN algorithm which overcomes the imbalance issue in Pima Indian diabetes dataset [6]. The SMOTE-ENN algorithm applies over-sampling with SMOTE to produce the constructed samples for minority imbalanced class, then applies cleaning techniques to under-sample with ENN to newly created instance [7]. Machine Learning ML is a field of computer science with various applications, including robotics, industry, education, enterprise, astronomy, and healthcare. Machine learning depends on learning from data by detecting underlying patterns and applying newly acquired knowledge to solve problems in previously unseen data [8].

There is clear interest in the application of machine learning and Artificial Intelligence AI to clinical research and practice. However, information on how to develop robust machine learning and AI approaches in medicine is insufficient [9]. Compared to the other ML methods, Support Vector Machines SVMs are very powerful in terms of recognizing patterns in complex data sets. However, the least square support vector machine is an improved algorithm of support vector machine [11]. In this study, we used the Least Square Support Vector Machine LS-SVM to perform classification tasks.

The performance of many machine learning algorithms is heavily dependent on the hyperparameters used. The hyperparameters such as the learning rate, kernel size, and number of estimators are usually set in many machine learning algorithms. They can be used to recognize handwriting or fraudulent credit cards, identify a speaker, and detect faces. The least square support vector machine is a powerful method for building a classifier. The Least square support vector machine is the improved algorithm of support vector machine [10][11]. The performance of many ML methods is heavily dependent on the hyperparameters used.

The hyperparameters such as the learning rate, kernel size, and number of estimators are usually set through investigation. They are determined before the learning process which estimates the optimized parameters of the model used [12]. Grid Search is a heuristic method and is used to find the finest possible values of parameters in a definite interval range to produce an optimal model. Grid Search is utilized in this work. It is an optimization approach which is conducted for the classification model parameter tuning that leads to the best performance [13]. Many performance metrics are currently available to evaluate the validity of machine learning algorithm in classification problems.

The proper interpretation of a performance metric that established on the problem domain and requirements is significant. Thus, the proposed algorithm figure 1, is evaluated using three performance metrics, they are accuracy, recall, and precision [14]. The remainder of this study is organized as follows. Section 2 provides the methodology and its analysis. Section 3 details the proposed algorithm and research work. Section 4 describes the dataset. Section 5 presents the experimental setup and discusses performance evaluation. The concluding remarks is provided in section 6.

2. METHODOLOGY

2.1 Pre-processing

Preprocessing is a crucial step prior to analyzing a dataset, it involves cleaning and modifying raw data to improve the information contained in the dataset. Data preprocessing is a demanding task, but it is necessary for putting data into context and reducing bias [15]. Outliers and missing values are often encountered during the data collection phase of experimental studies conducted in all fields of sciences and especially in the medical field, as it affects the treatment and diagnosis that the patient should receive. In addition, missing data in the medical dataset raises issues in the process of creating conclusion from case files [16]. Imputation is a method of handling missing values. Imputation replaces missing values with substituted values from a statistical analysis to produce a complete dataset without missing values for analysis. Imputation includes various methods, such as the mean, median, probability, ratio, regression, predictive regression, and assumption of distribution. The Pima Indian diabetes dataset has zero values. Thus, in the first preprocessing step of this study, the median value of the features with invalid zeros was imputed [17]. The second preprocessing step applied to the Pima Indian dataset is the resampling approach. Data resampling methods are generally used because of their simplicity and flexibility. The goal of resampling techniques is to rebalance the class

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

distribution of a dataset. These techniques are used to reduce the samples with low weights and increase the number of samples with high weights while retaining the total quantity of samples [18] [19].
The Synthetic Minority Over-Sampling Technique SMOTE is an over-sampling technique. Several samples are randomly selected for each minority class, sample x from their k nearest neighbors, and a new sample emerged as stated by equation (1). The new minority class instances can emerge. However, a problem which is sample overlap is generated since each minority class sample will produce a new sample.

$x_{new} = x_i + | x_i - x_i' | + \delta$ ... (1)

x_{new} is the new sample; x_i is the minority sample; x_i' is one of the k-nearest neighbors of x_i; δ is a random number and $\delta \in [0, 1]$.

The Edited Nearest Neighbor ENN technique is designed to identify and remove uncertain or noisy samples within a dataset. It operates by assessing each sample using the k-Nearest Neighbors k-NN rule against the rest of the data. If a sample belongs to the minority class and at least two of its three nearest neighbors belong to the majority class, the sample is removed. This process helps create smoother and more distinct boundaries between classes [20][21].

The SMOTE-ENN technique merges Synthetic Minority Oversampling Technique SMOTE with ENN. While SMOTE helps balance the data distribution by oversampling the minority class, it often cause issues like overlapping samples between the majority and minority classes, potentially limiting the performance of the classifier. To address this, SMOTE-ENN first oversamples the training data using SMOTE and then applies ENN to identify and remove misclassified samples by examining their three nearest neighbors. This results in cleaner and more well-defined data, improving classification accuracy [22][23][24]. In this study, SMOTE and ENN are used to enhance data quality and the proposed algorithm performance.

2.2 The Fine-Tuned Least Square Support Vector Machine LS-SVM

2.2.1 The Least Square Support Vector Machine LS-SVM

The Support Vector Machine SVM classifier was found more than a decade ago by (Vapnik, 1995). Support Vector Machines gained increasing attention because of its solid theoretical foundation. SVMs are a set of linked methods for supervised learning that are relevant for classification and regression problems [25] [26]. A modified algorithm of SVM proposed by Suykens and Vandewalle (1999), called Least Square Support Vector Machine LS-SVM [27]. The LS-SVM operates under equality instead of inequality constraints and uses the squared error cost function for judging the merit of algorithm optimization, giving it a remarkable advantage compared to SVM [28].

The SVM for classification, considers a binary classification training sample $\{(x_i, y_i)\}$ i=1, 2..., l, where x_i, is the vector of input pattern for the ith example and y_i, is the corresponding target output [29]. The pattern represented by the subset y_i=+1 belongs to class 1, and the pattern represented by the subset y_i =-1 belongs to class 2.

The original SVM classifier satisfies the following conditions:

$y_i[w^T q(x_i) + b] \geq 1$. i=1…l ….. (2)

The feature map is q: $R^n \rightarrow R^m$, that mapping the input space to a high dimensional feature space where the data points get linearly separable by a hyperplane specified by the pair ($w \in R^m$,$b \in R$).

Then the given classification function is,

$y(x) = \text{sign}\ \{w^T\ q(x) + b\}$ ….. (3)

It is unnecessary to compute with the feature map, and one only needs to work instead with a kernel function in the original space given by,

$K(x_i, x_j) = q(x_i)^T q(x_j)$ ….. (4)

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

In the case of noisy data, forcing zero training error will lead to poor generalization. To take account of the fact that some data points may be misclassified, introduce a set of slack variables:

$\xi_i \geq 0$, i=1…l

The relaxed separation constraint is given as,

$y_i[w^T q\ (x)_i + b] \geq 1 - \xi_i, \quad i = 1, ..., l$ ….. (5)

The optimal separating hyperplane can be found by the following minimization problem:

$$\underset{w,b,e}{Min} = J(w,b) = \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \qquad \text{…..(6)}$$

Subject to the constraints

$$y_i[w^T q\ (x)_i + b] \geq 1 - \xi_i \qquad i = 1, \dots, l$$
$$\xi_i \geq 0 \qquad i = 1, \dots, l$$

where *C* is a regularization parameter used to decide a tradeoff between the training error and the margin.

The Vapnik's standard SVM classifier formulation was modified by Suykens and Vandewalle into the following LS-SVM formulation:

$$\underset{w,b,e}{Min} = J(w,b) = \frac{1}{2} w^T w + \gamma\ \frac{1}{2}\sum_{i=1}^{l} e_i^2 \qquad \text{..... (7)}$$

subject to the equality constraint

$y_i[w^T q\ (x)_i + b] = 1 - e_i \quad i = 1, ..., l$ ..... (8)

we observe that the transition from equation 6 to equation 7 involves replacing the inequality constraints with equality constraints and adding a squared error term (forming a least squares), like to ridge regression. The corresponding Lagrange for equation 7 is:

$L(w,b,e,\alpha)) = J(w,e) - \sum_{i=1}^{l} \alpha_i\{y_i[w^T q\ (x)_i + b] - 1 + e_i\}$ …. (9)

where the α_i are Lagrange multipliers.

As was shown in ref 15, the optimality condition leads to the following

$(N + 1)\times(N + 1)$ linear system:

$$\begin{bmatrix} 0 & y^T \\ y & zz^T + y^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \text{..... (10)}$$

where $Z = [\ q\ (x_1)^T y_1; \ \dots\dots; q\ (x_l)^T y_l]$, $Y = [y_1; \dots\dots; y_l]$, $1 = [1...; 1]$. ..... (11)

Mercer's condition is applied within the matrix zz^T:

$zz^T = y_i y_j\ q\ (x_i)^T q\ (x_j) = y_i y_j K\ (x_i, x_j)$ ..... (12)

Thus, we would only need to use kernel function *K* in the training algorithm and would never need to explicitly even know what q *is*.

The LS-SVM classifier is then constructed as follows:

$f(x) = sign\ (\sum_{i=1}^{l} y_i \alpha_i\ K\ (X, X_i) + b)$ ..... (13)

2.2.2 Hyperparameter optimization

The hyperparameters optimization that is also referred to as model selection, is the problem of choosing a set of hyperparameters for a model. The goal is to optimize the performance metrics of a model an independent dataset and ensure that the model does not overfit its data by tuning [30][31]. Grid Search is the most frequently used hyper-

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

parameter optimization technique. It performs exhaustive search on a manually specified subset of hyper-parameter Space. The GridSearchCV is an algorithm for detecting the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made [32][33]. The gride search performs an exhaustive search on a manually specified subset of the hyperparameter space.

Let X be a target algorithm with k parameters to be tuned, and let parameter θ_i be a value within the interval $[x_i, y_i\ ]$ in the parameter search space:

$$\Theta = [x_1, y_1\ ] \times \ldots\ldots \times [x_k, y_k].$$

Θ: H → R is a performance measurement function that maps θ to a numeric score.

A grid search attempts to evaluate every combination of hyperparameters and note the accuracy. When all combinations are evaluated, the model provides the set of parameters with the best accuracy

Cross-validation, which is a splitting strategy used to evaluate the fitness of the parameter values in a grid search to optimize parameters such as C and gamma in an LS-SVM classifier [35][36]. LS-SVM was trained with different coefficients that were determined through the optimization algorithm. The trained fine-tuned LS-SVM model was tested using validation diabetes data [37] [38].

3. MODEL PERFORMANCE EVALUTION

Many performance metrics exist to evaluate the LS-SVM classifier algorithm. Computing the number of correctly detected class samples (true positives), the correctness of a classification estimated. The number of correctly detected samples that do not belong to the class (true negatives), and samples that either were incorrectly ascribed to the class (false positives) or that were not detected as class samples (false negatives). The most utilized classification measures are the precision and recall metrics that are used in information retrieval. The classification accuracy is in terms of percentage.

The higher the values of precision and recall are, the much better the classifier. Accuracy A is the overall effectiveness of a classifier. It is the proportion of the total number of predictions that are correct, and it highly depends on the dataset distribution to determine the system performance. A model is said to be satisfactory when the accuracy is high. The Recall R is effectiveness of a classifier to identify positive labels, and its proportion of relevant subjects who are correctly identified. The Precision P is the proportion of predicted relevant subjects who are correctly identified [39] [40].

The validation measurements used in this study are as follows:

$$R = \frac{TP}{FN+TP} \qquad \ldots\ldots (14)$$

$$P = \frac{TP}{FP+TP} \qquad \ldots\ldots (15)$$

$$A = \frac{TN+TP}{TN+FN+TP+FP} \qquad \ldots\ldots (16)$$

where

TP is the true positives (i.e., patients correctly classified),

FN is the false-negatives (i.e., patients incorrectly classified),

TN is true negatives (i.e., patients not related to the condition and classified correctly), and

FP is a false-positives (i.e., patients related to the condition and classified incorrectly).

4. THE DATA SETS DISCRIPTION

Pima Indian diabetes dataset is a collection of data from subjects with and without Type 2 Diabetes T2D. The subjects in this dataset are females of Pima Indian heritage who are at least 21 years old [41]. The Pima Indian

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

diabetes dataset was accessed from the Kaggle data depot. The source data for this dataset was provided by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset was used to diagnose whether a patient has diabetes based on certain diagnostic measures. The data are numerical and contain a total of 8 features as listed in Table 1 and 768 samples [42].

The Iraqi Society Diabetes ISD dataset was accessed from the Mendeley that gathered from the Iraqi society. The dataset was acquired from the laboratory of Medical City Hospital and the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital. The patients' files were collected, and data produced from the files and entered to the database to build the diabetes dataset. The dataset consists of 1000 patients and has three classes (Diabetic, Non-Diabetic, and Predicted- Diabetic) and 14 attributes as listed in Table 2. We excluded the Predict-Diabetic class since the group is small [43].

TABLE I. Pima dataset attributes

No	Attribute	Data Type	Note
1	Preg	Numerical	The number of pregnancies
2	Gluc	Numerical	Glucose plasma levels two hours after consuming glucose
3	Bp	Numerical	Diastolic blood pressure (mm Hg)
4	Skin	Numerical	Thickness of the skin fold on the triceps of the upper arm (mm)
5	Insulin	Numerical	Insulin serum levels in the blood two hours after the glucose test (lh/ml)
6	BMI	Numerical	Body mass index [weight in kg/(Height in m)], an index used to evaluate a person's relative weigh
7	DPF	Numerical	Diabetes pedigree function is a value that measures genetic risk factors based on a family history of diabetes
8	Age	Numerical	Patient's age in years

TABLE II. The ISD Dataset attributes

No	Attribute	Data Type	Note
1	ID	Numerical	Identity of patient
2	No-Patient	Numerical	No. of patients
3	Gender	Categorical	Gender
4	Age	Numerical	Age (years)
5	Urea	Numerical	Urea
6	Cr	Numerical	Creatinine ratio
7	HbA1C	Numerical	Hemoglobin A1C
8	Chol	Numerical	Cholesterol
9	TG	Numerical	Triglycerides
10	HDL	Numerical	High-density lipoprotein
11	LDL	Numerical	Low-density lipoprotein
12	VLDL	Numerical	Very low-density lipoprotein
13	BMI	Numerical	Body mass index (weight in kg/(height in m)2)

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

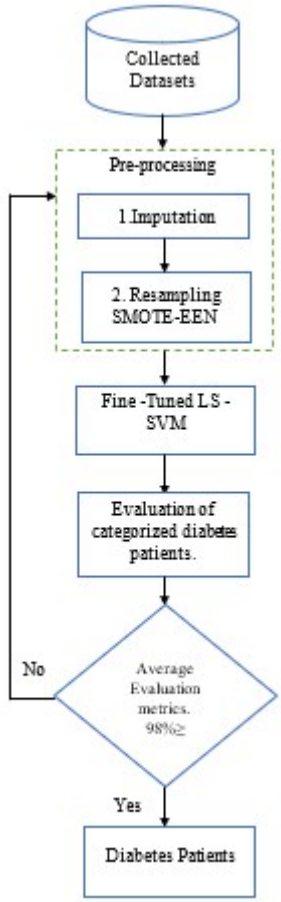

Fig.1 The proposed SMOTE-ENN with fine-tuned LS-SVM algorithm

5. RESULTS AND DISCUSSION

The results for the proposed algorithm carried out in this work are presented in this section.Tests with real-world datasets were evaluated. We used PIMA Indians and ISD diabetes datasets. The datasets are preprocessed firstly using imputation technique. The LS-SVM algorithm applied to Pima Indian diabetes dataset. The registered results were not satisfied, the accuracy measurement was 75%. The SMOTE-ENN is used for resampling imbalanced Pima and ISD datasets. The resampling technique has significant effect on LS-SVM algorithm performance, it enhanced the metrics results, in which the accuracy measurement increased, and it was 91%. The LS-SVM algorithm is fine-tuned with GridSearchCV algorithm that optimizes the hyperparameters C and gamma. The fined tuned LS-SVM with resampling technique SMOTE-ENN showed efficient classification results as shown in table 3, it is 98%. Table 3 provides the performance metric values for Pima dataset that are obtained from proposed efficient classification algorithm.

TABLE III: Results of the Least Square Support Vector Machine algorithm and fine-tuned Least Square Support Vector Machine algorithm with SMOTE-ENN applied to the Pima dataset

Algorithm	Accuracy	Precision	Recall
LS-SVM	75%	70%	55%
SMOTE-ENN-LS-SVM	91%	92%	93%
SMOTE-ENN-Fine-tuned LS-SVM(Our)	98%	98.%	97%

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

The Least Square Support Vector Machine algorithm applied to the ISD dataset, the LS-SVM algorithm showed accepted results, the accuracy measurement was 97.3 %. The ISD diabetes dataset is preprocessed and resampled with SMOTE-ENN. The result as shown in table 4 is enhanced. The registered accuracy was 98%. The LS-SVM algorithm is optimized with GridSearchCV algorithm, thus efficient classification results are optioned, and the accuracy was 99.8%. Table 4 provides the performance metrics values from the SMOTE-ENN with fine-tuned LS-SVM algorithm for ISD dataset.

Table IV: Results of the Least Square Support Vector Machine algorithm and fine-tuned Least Square Support Vector Machine algorithm with SMOTE-ENN applied to the Iraqi Society Diabetes ISD.

Algorithm	Accuracy	Precision	Recall
LS-SVM	97.3%	97.9%	99%
SMOTE-ENN-LS-SVM	98%	98%	99%
SMOTE-ENN-Fine-tuned LS-SVM	99.8%	99.7%	99.8%

6. CONCLUSIONS

In this study, to develop a classification model for the outcomes of the clinical Pima Indian diabetes and Iraqi Society Diabetes ISD datasets, we proposed a classification algorithm as in figure 1, that is SMOTE-ENN with fine-tuned LS-SVM algorithm. The LS-SVM is applied on both datasets. The LS-SVM algorithm optimized using GridsearchCV, and it improved the LS-SVM algorithm performance by optimizing the C and gamma hyperparameters. In addition, the fine-tuned LS-SVM performance was enhanced when the SMOTE-ENN resampling technique was used as the preprocessing step and the median imputation. We observed that the proposed algorithm obtained the highest average values of the evaluation metrics, the accuracy obtained values were 98% and 99.8%, recall values were 97%and 99.8% , and precision values were 97% and 99.7%.

The data availability statement

The data sets supporting this study are freely available in [Pima Indians Diabetes Database, at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database] and the Iraqi Society Diabetes ISD dataset,[https://data.mendeley.com/datasets/wj9rwkp9c2/1].

Funding

The authors received no financial support for this research.

References

[1] R. Knevel and K. P. Liao, "From real-world electronic health record data to real- world results using artificial intelligence," Annals of the Rheumatic Diseases, vol. 82, pp. 306-311, 2023.

[2] J.-H. Lin and P. J. Haug, "Data preparation framework for preprocessing clinical data in data mining," in AMIA annual symposium proceedings, 2006, p. 489.

[3] I. Battas, R. Oulhiq, H. Behja, and L. Deshayes, "A Proposed Data Preprocessing Method for an Industrial Prediction Process," in 2020 6th IEEE Congress on Information Science and Technology CiSt, 2021, pp. 98-103.

[4] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," Journal of Big Data, vol. 8, pp. 1-37, 2021.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[5] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," GESTS international transactions on computer science and engineering, vol. 30, pp. 25-36, 2006.

[6] N. V. Chawla, "Data mining for imbalanced datasets: An overview," Data mining and knowledge discovery handbook, pp. 875-886, 2010.

[7] D. T. Ludera, "Credit card fraud detection by combining synthetic minority oversampling and edited nearest neighbors," in Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference FICC, Volume 2, 2021, pp. 735-743.

[8] A. Tuppad and S. D. Patil, "Machine learning for diabetes clinical decision support: a review," Advances in Computational Intelligence, vol. 2, p. 22, 2022.

[9] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," BMC Medical Research Methodology, vol. 22, pp. 1-15, 2022.

[10] W. Sun and C. Yang, "Research of least square support vector machine based on chaotic time series in power load forecasting model," in Neural Information Processing: 13th International Conference, ICONIP 2006, Hong Kong, China, October 3-6, 2006. Proceedings, Part II 13, 2006, pp. 984-993.

[11] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine SVM learning in cancer genomics," Cancer genomics & proteomics, vol. 15, pp. 41-51, 2018.

[12] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization," Journal of Information Security and Applications, vol. 55, p. 102596, 2020.

[13] A. R. Laeli, Z. Rustam, S. Hartini, F. Maulidina, and J. E. Aurelia, "Hyperparameter Optimization on Support Vector Machine using Grid Search for Classifying Thalassemia Data," in 2020 International Conference on Decision Aid Sciences and Application DASA, 2020, pp. 817-821.

[14] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza, "General Performance Score for classification problems," Applied Intelligence, vol. 52, pp. 12049-12063, 2022.

[15] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, "Data Preprocessing," in Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics, ed: Springer, 2022, pp. 117-163.

[16] M. F. Dzulkalnine and R. Sallehuddin, "Missing data imputation with fuzzy feature selection for diabetes dataset," SN Applied Sciences, vol. 1, p. 362, 2019.

[17] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," Korean journal of anesthesiology, vol. 70, pp. 407-411, 2017.

[18] J. Li, Y. Wu, S. Fong, A. J. Tallón-Ballesteros, X.-s. Yang, S. Mohammed, et al., "A binary PSO-based ensemble under-sampling model for rebalancing imbalanced training data," The Journal of Supercomputing, pp. 1-36, 2022.

[19] T. Lu, Y. Huang, W. Zhao, and J. Zhang, "The metering automation system based intrusion detection using random forest classifier with smote+ enn," in 2019 IEEE 7th International Conference on Computer Science and Network Technology ICCSNT, 2019, pp. 370-374.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[20] A. Channa, O. Cramariuc, M. Memon, N. Popescu, N. Mammone, and G. Ruggeri, "Parkinson's disease resting tremor severity classification using machine learning with resampling techniques," Frontiers in Neuroscience, vol. 16, p. 955464, 2022.

[21] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, pp. 273-297, 1995.

[22] A. Shmilovici, "Support vector machines," Data mining and knowledge discovery handbook, pp. 257-276, 2005.

[23] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural processing letters, vol. 9, pp. 293-300, 1999.

[24] H. Zheng, J. Liu, R. Zhuang, F. Zhao, M. Zhen, Y. Wang, et al., "Enhancing the performance of LSSVM model in predicting rock fragmentation size via optimization algorithms," Ksce Journal of Civil Engineering, vol. 27, pp. 3765-3777, 2023.

[25] X. Yao, H. Liu, R. Zhang, M. Liu, Z. Hu, A. Panaye, et al., "QSAR and classification study of 1, 4-dihydropyridine calcium channel antagonists based on least squares support vector machines," Molecular pharmaceutics, vol. 2, pp. 348-356, 2005.

[26] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," BMC Medical Research Methodology, vol. 22, pp. 1-15, 2022.

[27] G2 G. O. Anyanwu, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "Optimization of RBF-SVM Kernel using Grid Search Algorithm for DDoS Attack Detection in SDN-based VANET," IEEE Internet of Things Journal, 2022.

[28] S. Pattanayak and T. Singh, "Cardiovascular disease classification based on machine learning algorithms using gridsearchcv, cross validation and stacked ensemble methods," in International Conference on Advances in Computing and Data Sciences, 2022, pp. 219-230.

[29] D. J. Kalita, V. P. Singh, and V. Kumar, "A survey on SVM hyper-parameters optimization techniques," in Social Networking and Computational Intelligence: Proceedings of SCI-2018, 2020, pp. 243-256.

[30] G. Behera and N. Nain, "GSO-CRS: grid search optimization for collaborative recommendation system," Sādhanā, vol. 47, p. 158, 2022.

[31] A. R. Laeli, Z. Rustam, S. Hartini, F. Maulidina, and J. E. Aurelia, "Hyperparameter Optimization on Support Vector Machine using Grid Search for Classifying Thalassemia Data," in 2020 International Conference on Decision Aid Sciences and Application DASA, 2020, pp. 817-82.

[32] S. Pattanayak and T. Singh, "Cardiovascular disease classification based on machine learning algorithms using gridsearchcv, cross validation and stacked ensemble methods," in International Conference on Advances in Computing and Data Sciences, 2022, pp. 219-230.

[33] D. J. Kalita, V. P. Singh, and V. Kumar, "A survey on SVM hyper-parameters optimization techniques," in Social Networking and Computational Intelligence: Proceedings of SCI-2018, 2020, pp. 243-256.

[34] G. Behera and N. Nain, "GSO-CRS: grid search optimization for collaborative recommendation system," Sādhanā, vol. 47, p. 158, 2022.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 119-129
DOI: https://doi.org/10.25195/ijci.v51i1.579
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[35] G. Ranjan, A. K. Verma, and S. Radhika, "K-nearest neighbors and grid search cv based real time fault monitoring system for industries," in 2019 IEEE 5th international conference for convergence in technology I2CT, 2019, pp. 1-5.

[36] M. S. Borujeni, M. Ghaderi-Zefrehei, F. Ghanegolmohammadi, and S. Ansari-Mahyari, "A novel LSSVM based algorithm to increase accuracy of bacterial growth modeling," Iranian Journal of Biotechnology, vol. 16, 2018.

[37] B. Habib and F. Khursheed, "Performance evaluation of machine learning models for distributed denial of service attack detection using improved feature selection and hyper-parameter optimization techniques," Concurrency and Computation: Practice and Experience, vol. 34, p. e7299, 2022.

[38] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information processing & management, vol. 45, pp. 427-437, 2009.

[39] P. Sankar Ganesh and P. Sripriya, "A comparative review of prediction methods for pima Indians diabetes dataset," Computational Vision and Bio-Inspired Computing: ICCVBIC 2019, pp. 735-750, 2020.

[40] Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on Machine Learning ML algorithms," Neural Computing and Applications, pp. 1-17, 2022.

[41] M. F. Aslan and K. Sabanci, & quot; A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data, & quot; Diagnostics, vol. 13, p. 796,2023.

[42] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and feature analysis for Diabetes Mellitus classification using random forest," Journal of Computing Theories and Applications JCTA, vol. 1, pp. 41-49, 2023.

[43] X. Li, M. Curiger, R. Dornberger, and T. Hanne, "Optimized computational diabetes prediction with feature selection algorithms," in Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, 2023, pp. 36-43.