IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.18, No.1, January 2024, pp. 1~12
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.91857

Detection of Hospital Claim Anomalies Using Support Vector Regression

Luthfia Nurma Hapsari*[1], Nur Rokhman[2]
[1]Master Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]luthfianurmahapsari@mail.ugm.ac.id**, [2]nurrokhman@ugm.ac.id

Abstract

BPJS Kesehatan berperan penting terhadap akses terjangkau layanan kesehatan dan mengurangi beban finansial perorangan. Namun demikian, masalah defisit dapat mengganggu keberlanjutan program. Oleh karena itu deteksi anomali sangat penting untuk dilakukan.

Berdasarkan Peraturan Menteri Kesehatan No 52 Tahun 2016 (Permenkes), terdapat 7 variabel independen yang memengaruhi nilai klaim Group Tarif Indonesian Case Base Groups (Group Tarif INACBGs) yang menentukan biaya yang dibayarkan BPJS Kesehatan kepada Rumah Sakit. Hubungan antar variable ini bisa bersifat linear atau non-linear kompleks. Oleh karena itu, digunakan Regresi Linear Berganda (RLB) dan Support Vector Regression (SVR) kernel Radial Basis Function (RBF) untuk deteksi anomali. Hasil deteksi anomali dari keduanya dibandingkan untuk menentukan algoritme terbaik.

*Penelitian menunjukkan bahwa SVR RBF mengungguli RLB dalam deteksi anomali. SVR RBF menghasilkan Accuracy=0,97, Precision=0,84, Recall=0,97, dan F1-Score=0,90 dengan parameter C=1, epsilon=1000, gamma=1000, dan definisi anomali > 0,5 * RMSE pada Dataset Normalization serta Dataset PCA. Model SVR RBF yang dilatih dengan Dataset PCA menonjol dalam kecepatan waktu eksekusi dan memberikan hasil deteksi anomali sebanding dengan Dataset Normalization.*

Keywords*—Anomaly Detection, BPJS Kesehatan, Support Vector Regression, PCA*

Abstract

BPJS Kesehatan plays an important role in providing affordable access to healthcare services and reducing individual financial burdens. However, deficit issues may disrupt the sustainability of the program, making anomaly detection highly important to conduct.

Based on Minister of Health Regulation No. 52/2016 (Permenkes), there are seven independent variables that affect the value of Indonesian Case Base Groups Tariff claims (INACBGs Tariff Group), determining the fees paid by BPJS Kesehatan to hospitals. This relationship can be linear or complex non-linear. Therefore, Multiple Linear Regression (MLR) and Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel are used. Anomaly detection results from both methods were compared to determine the best algorithm.

*The research shows that SVR RBF outperforms MLR in anomaly detection. SVR RBF resulting in Accuracy=0,97, Precision=0,84, Recall=0,97, and F1-Score=0,90 with parameters C=1, epsilon=1000, gamma=1000, and anomaly definition >0,5*RMSE on the Normalization Dataset and PCA Dataset. The SVR RBF model trained with the PCA Dataset stands out in execution time speed and provides comparable anomaly detection results to the Normalization Dataset.*

Keywords— *Anomaly Detection, BPJS Kesehatan, Support Vector Regression, PCA*

Received December 13th,2023; Revised January 22th, 2024; Accepted January 29th, 2024

1. INTRODUCTION

BPJS Kesehatan, as Indonesia's Social Health Insurance Organizer, plays an important role in enhancing healthcare access and quality for Indonesian citizens. It aims to provide affordable health services, alleviate individual financial burdens, and ensure the financial stability of healthcare providers [1] .

The financial burden on BPJS Kesehatan, exacerbated by fraudulent claims and program misuse, contributes to the deficit in the National Health Insurance Program (JKN) [2], [3]. Detecting anomalies in healthcare data is needed to prevent fraud, ensuring program integrity, and maintaining financial sustainability [4].

Previous research focused on unsupervised anomaly detection using Simple Linear Regression (SLR) for anomaly detection in the value of verified fee claims paid by BPJS Kesehatan to hospitals (BPJS Verified Fees). The BPJS Kesehatan dataset 2015-2016 which considered two independent variables, namely length of stay and hospital bill costs to BPJS Kesehatan (Hospital Bill Fees) was modeled with SLR to detect anomalies in BPJS Verified Cost claim values [5], [6]. But, the performance evaluation of the model was not carried out due to the absence of ground truth in previous studies.

This research proposes a different approach. This research proposes the detection of supervised anomalies in the claim value of Indonesian Case Base Groups (INACBGs Tariff Group) in the BPJS Kesehatan dataset 2015-2018. The INACBGS Tariff Group covers all components of hospital resources, determines Hospital Bill Fees and BPJS Verified Fees. The INACBGs Tariff Group is determined based on Minister of Health Regulation No. 52/2016 (Permenkes), which plays as the ground truth of this study.

Based on Permenkes, there are seven variables in the form of numerical and categorical data that affect the claim value of the INACBGs Tariff Group. These variables consist of severity, participant treatment class, INACBGs Casemix Main Groups (CMG) Code, INACBGs regional rates, Advanced Level Referral Health Facilities BPJS Kesehatan (FKRTL), FKRTL Type, and FKRTL Service Level.

In this study, two regression algorithms were chosen to detect anomalies for seven independent variables that affect the claim value of the INACBGs Tariff Group. The relationship between the independent variable and the target can be linear or nonlinear complex. Multiple Linear Regression (MLR) and Support Vector Regression (SVR) of the Radial Basis Function (RBF) kernel were chosen. MLR is suitable for accommodating linear relationships between two or more independent variables and target variables. SVR RBF is suitable for accommodating complex nonlinear relationships between independent variables and targets [7], [8]. SVR RBF was chosen based on its ability to accommodate non-linear relationships efficiently, particularly in the context of various independent variables affecting the target variable claim value of the INACBGs Tariff Group. The RBF kernel's flexibility in capturing complex data patterns and its performance in various domains make SVR as a reliable choice [9], [10]. The use of RBF in SVR is crucial for handling non-linear relationships. This will be explicitly described in the Methods section.

The MLR and SVR RBF performance in anomaly detection will be compared after initial testing with a normalized dataset (Normalization Dataset). SVR RBF algorithms that provide better results in anomaly detection will be hyperparameterized, and the performance will be compared to datasets with dimension reduction using Principal Component Analysis (PCA Dataset). The goal is to improve the identification of anomalies more optimally and quickly, and contribute to the sustainability of the BPJS Kesehatan program in Indonesia.

The Normalization Dataset and PCA Dataset refer to the result of specific preprocessing techniques. Further explanation is presented in the Methods section to provide a clear understanding on the dataset preparation.

2. METHODS

There are four main processes in this research, namely: data preparation, building MLR model with Normalization Dataset, building SVR model with Normalization Dataset, and building SVR model with PCA Dataset.

The data preparation stage includes random sampling of FKRTL data, preprocessing of FKRTL data, and processing of Permenkes data. The preprocessing stage consists of feature selection, standardization, codification, normalization, and dimension reduction with Principal Component Analysis (PCA).

The MLR model was trained and tested with the Normalization Dataset, as shown in Figure 1. The SVR RBF model was trained and tested with the Normalization Dataset. Tuning on SVR RBF hyperparameter and anomaly definition was carried out in the SVR Normalization Model. This produced the best combination of Accuracy, Precision, Recall, and F1-Score. The SVR Normalization model was validated with K-Fold cross validation. This third main process is shown in Figure 2.

Hyperparameters and anomaly definitions that produce the best results in previous process was used to build SVR RBF model with the PCA Dataset. The SVR model was trained, tested and validated with the PCA Dataset. This fourth main process is shown in Figure 3.

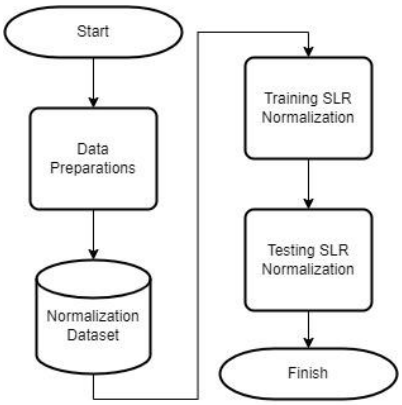

Figure 1 Flowchart of Building MLR Models with Normalization Dataset

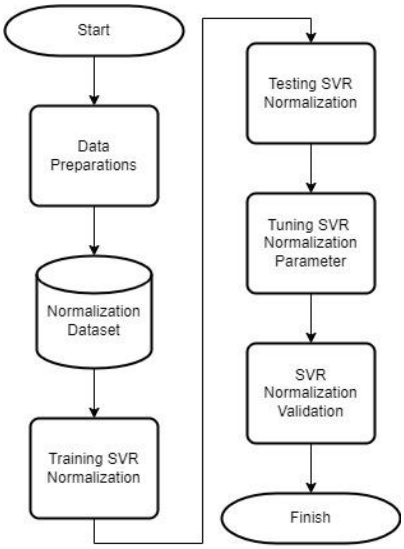

Figure 2 Flowchart of Building SVR RBF Models with Normalization Dataset

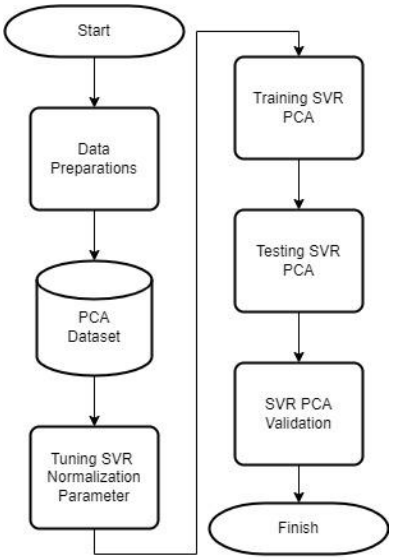

Figure 3 Flowchart of Building SVR RBF Models with PCA Dataset

2.1 Implementation Tools

This research used an Intel(R) Core(TM) i5-8250U CPU @ 1,60GHz 1,80 GHz processor and 8,00 GB (7,86 GB usable) RAM with operating system of Windows 10, Python 3.10.12 as the programming language, and tools such as Google Colab and Jupyter Notebook as integrated development environments (IDEs). The essential libraries utilized in the study involve pandas, sklearn, google.colab, matplotlib, seaborn, datetime, scipy, numpy, joblib, os, tabula, time, and unicodedata.

2.2 Random Sampling of FKRTL

The BPJS Kesehatan data that used in this study was the FKRTL Dataset. Generally, fraud is more common in advanced healthcare facilities such as hospitals, on the otherhand fraud is less common in primary health facilities such as clinics [11]. FKRTL data includes data on hospital entitlements, medical expenses, and patient illnesses. Health insurance fraud can be detected by data anomalies resulting from a combination of measures and diagnostics [12].

The FKRTL dataset contains 2.509.743 rows of data. Random sampling was carried out for 100.000 rows of data. From 100.000 rows were devided into 70% or 70.000 rows as the model training and 30% or 30.000 rows as the model testing. The sample taken has 3 conditions, namely: (1) data with five Special Tariff variables are empty, null, or zero, (2) data with the variable INACBGs Tariff Group (tariff_group) equal to the Hospital Bill Fees (billing_costs) or equal to the BPJS Verified Fees (verified_costs), (3) data with no missing values in the 14 variables used in the Preprocessing – Feature Selection stage.

2.3 Preprocessing of FKRTL

Preprocessing needs to be carried out to make it ready to train and test SVR and MLR models.

2.3.1 Feature Selection

There are 55 columns or variables in the FKRTL data. Fourteen variables was selected, namely: visit_ID (FKL02), arrival_date (FKL03), discharge_date (FKL04), health_facility_ownership (FKL07), health_facility_type (FKL09), service_level (FKL10), treatment_class (FKL13), inacbgs_code (FKL19), cmg (FKL20), severity_level (FKL23), regional (FKL31), INACBGs Group Tariff / tariff_group (FKL32), billing_costs (FKL47),

verified_costs (FKL48). These variables are considered to affect the value of tariff_group claims or used in the next processes while other variables only contain demographic data of patients and health facilities.

2.3.2 Standardization

Data standardization is the process of converting data to a common format to allow users to process and analyze it. Standardization carried out at FKRTL included converting text to lowercase, changing to int data type, and changing to string data type.

2.3.3 Codefication

The codefication process converted the categorical data into numeric data. One method commonly used in the codification process is Label Encoding which converts categorical data to the nearest integer format [13]. Codefication was carried out on seven variables, namely severity_level, treatment_class, cmg, regional, health_facility_ownership, health_facility_type, and service_level by using Label Encoding method.

2.3.4 Normalization

Normalization was carried out to ensure that all variables have the same range using the Z-score [13]. The general equation for Z-score is shown in Equation 1.

$$Z(x_{ij}) = \frac{x_{ij} - x_j}{a} \quad (1)$$

x_j is the average of the data, x_{ij} is the original value and a is the standard deviation. All variables were normalized. Normalization was carried out on seven variables which influence the claim value of the INACBG Tariff Group (tariff_group) based on Permenkes. These seven variabels are severity_level, treatment_class, cmg, regional, health_facility_ownership, health_facility_type, and service_level. The normalization produced normalized sample dataset. This called the Normalization Dataset.

2.3.5 Principal Component Analysis (PCA)

The PCA is a method for reducing features from high to low dimensions by retaining as much information from the original dataset as possible [13]. Finding the PCA value requires computating the value of the covariance matrix and finding the eigenvalue and eigenvector [13]. PCA was carried out on seven variables which influence the claim value of the tariff_group based on Permenkes, namely: severity_level, treatment_class, cmg, regional, health_facility_ownership, health_facility_type, and service_level. The PCA process resulted three main components PC1, PC2, and PC3 features. The dimension of Normalization Dataset was dimensionally reduced by using PCA. This called the PCA Dataset.

2.4 Building the ground truth data

The Minister of Health Regulation No. 52/2016 (Permenkes) is in the form of document. Numerical data in the form of Excel or CSV was needed. Then, it can be used as the ground truth to determine whether the variables tariff_group in the FKRTL BPJS Kesehatan in accordance with the provisions of Permenkes or not. If it is appropriate, it will be labeled as a normal data, otherwise it will be labeled as an anomaly data. The result of this process was the Ground Truth Sample Dataset. It was used as a reference to evaluate the performance of SVR and MLR algorithms in detecting the anomalies.

2.5 Multiple Linear Regression (MLR)

MLR is a statistical method used to model the linear relationship between two or more independent variables and the target variable [14]. The general equation for MLR is shown in Equation 2.

$$[Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon] \quad (2)$$

where Y is the dependent variable, $(X_1, X_2, \ldots, X_n)$ is the independent variable, β_0 is the intercept, $(\beta_1, \beta_2, \ldots, \beta_n)$ is the regression coefficient (slope), ε is a random error.

2.6 Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel

SVR is a forecasting method that has good nonlinear prediction performance using Kernel functions which suitable for high-dimensional data sets. SVR uses kernel functions to map data into higher dimensions of feature space such that it can handle non-linear relationships between predictor variables and target variables [15]. The general equation for SVR is shown in Equation 3.

$$y = b_0 + \sum \alpha_i \cdot K(X_i, X) + b \quad (3)$$

where y is the dependent variable (target variable) to be predicted. X_i is a training data vector. X is the test data vector to be predicted. α_i is the coefficient obtained from the learning process. K is a kernel function used to map data into a higher feature space. b_0 and b are the constants (intercepts) of the model.

This research used RBF kernel that uses Gaussian functions to map data into a high-dimensional feature space, which suitable for modeling complex non-linear relationships [15]. The general equation for RBF is shown in Equation 4.

$$K(x_i, x_j) = e^{-\frac{|x_i - x_j|^2}{2\sigma^2}} \quad (4)$$

where (x_i) and (x_j) are the input vector. (e) is the basis of exponential. $K(x, x_i)$ is the result of RBF kernel functions. (σ) is the parameter that controls how fast the RBF kernel value degenerate as the distance between (x_i) and (x_j) increases.

2.7 Anomaly Detection

SVR RBF and MLR performance evaluations were carried out at the training, testing, and validation stages. SVR RBF and MLR performance evaluation for anomaly detection was measured by using Accuracy, Precision, Recall, and F1-Score. Accuracy measures the extent to which the model can correctly classify instances that belong to anomalies and those are not anomalies. Precision measures the extent to which instances classified as anomalies by the model are truly anomalies. Recall measures the extent to which the model can find all instances that are actually anomalies. F1-Score combines precision and recall to provide a balanced measurement between the two metrics. In the context of anomaly detection, these metrics evaluate how well the model can identify instances that are actually anomalies (True Positives) without errors (False Positives). The success of anomaly detection was measured by the balance between the three metrics.

Detecting anomalies in both algorithms at the training and testing stages consisted of the following steps. First, the SVR model and MLR model used to predict the target variables of each model. Second, calculating the residual, the difference between the actual value of the target variable and the predicted value of the target variable, as shown in Equations 5. A big number of

residue indicates a significant difference between the model's predictions and the actual value of the target variable [8].

$$Residue_i = y_i - \hat{y}_i \tag{5}$$

where $Residue_i$ is the residue for the (i)-th observation, y_i is the actual value of the target variable for the (i)-th observation. y_i is the predicted result of the regression model for the (i) observation. Third, the calculation of the RMSE. RMSE is defined as the square root of the residual. Fourth, determine the anomaly definition. The anomaly definition is the threshold of a data categorized as normal data or anomalous data based on its residue [10], as shown in Equation 6.

$$Anomali\ Definition_i = \text{Residue}_i > k \times RMSE \tag{6}$$

The $Anomali\ Definition_i$ is a variable that indicates whether the (i)-th observation is considered an anomaly based on the general definition [16]. $Residue_i$ is the residual value for the (i)-th observation. The $Residue_i$ is the difference between the actual value of the target variable and the predicted result of the regression model. The *k* value is the threshold used to determine wether the $Residue_i$ is considered a normal or an anomaly. The *k* value can be determined according to the characteristics of the data. In the definition of a general anomaly the value of *k* is 2 or 3. If the $Residue_i$ exceeds 2 or 3 times the RMSE, then the (i)-th observation is considered an anomaly [17]. Fifth, the performance evaluation of anomaly detection in both algorithms by comparing the results of anomaly detection with the Ground Truth Sample Dataset.

2.8 Comparison of SVR RBF and MLR Models in Normalization Datasets

At this stage, the anomaly detection performance of MLR and SVR RBF was compared to determine the best performance algorithm in terms of Accuracy, Precision, Recall and F1-Score. The dataset processed at this stage was Normalization Dataset. The Normalization dataset consist of 100.000 data were divided into 70.000 training data and 30.000 testing data.

Both the training stage and the testing stage of SVR and MLR models required independent variables and target variables. The independent variables and targets of the Normalization Dataset are presented in Table 1.

Table 1. Independent and Target Variables Dataset Normalization in the Training and Testing Stages of SVR and MLR Algorithms

Algorithm	**Independent Variables**	**Number of Vars. Independent**	**Target Variable**
SVR & MLR	'severity_level', 'treatment_class', 'cmg', 'regional', 'health_facility_ownership', 'health_facility_type', 'service_level'	7	'tariff_group'

The performance evaluation of SVR and MLR models in detecting anomalies used Accuracy, Precision, Recall, and F1-Score. These metrics were used to determine how well SVR models detect anomalies. An anomaly is detected if the residue greater than 2 * RMSE reference [5]. Data rows that have a residue greater than 2 * RMSE are detected as anomalies, while data less than or equal to 2 * RMSE are detected as normal.

At the training and testing stages for both algorithms, hyperparameter tuning was not carried out (using default hyperparameters) to see the performance of the algorithm with default hyperparameters. The SVR and MLR performance evaluation was compared to get the best model for anomaly detection.

2.9 Tuning SVR RBF Hyperparameters & Anomaly Definition on Normalization Dataset

SVR RBF hyperparameters tuning and anomaly definition were carried out to adjust the parameters needed to create the SVR model for anomaly detection. SVR RBF hyperparameters include C parameters to control the trade-off between error penalties and model complexity, gamma for non-linear kernels, and epsilon (to set fault tolerance). Two sets of hyperparameter values were applied for C, epsilon, and gamma, specifically, big values of 1000 and small values of 1.

The anomaly was detected when the residue or difference between the actual target variable and the predicted target variable greater than *k* x RMSE. The value of *k* was tuned to get the optimal Accuracy, Precision, Recall, and F1-Score. Two *k* values were used, namely 2 and 0,5.

2.10 Comparison of SVR Normalization Model with SVR PCA Model

At this stage, the PCA Dataset was used. The PCA dataset contained 100.000 records. Then, it was divided into 70.000 training data and 30.000 testing data. The SVR hyperparameters and the best anomaly definition tuning results at the stages: training, testing, and validation of SVR models was applied. The evaluation of model performance in anomaly detection was carried out. Training, testing, and validation stages of the SVR model require independent variables and predicted target variables. The independent variables and targets of the PCA Dataset and Normalization Dataset are presented in Table 2.

Table 2. Independent & Target Variables on Normalization Dataset & PCA Dataset in the Training, Testing, & Validation Stages of SVR RBF Algorithm

Dataset	**Independent Variables**	**Number of Independent Vars.**	**Target Variable**
Normalization	'severity_level', 'treatment_class', 'cmg', 'regional', 'health_facility_ownership', 'health_facility_type', 'service_level'	7	'tariff_group'
PCA	'PC1','PC2','PC3'	3	'tariff_group'

Appliying the PCA to the Normalization Dataset resulted three principal components (PC1, PC2, and PC3). PC1, as the primary component, captured the highest variability in the data, representing the dominant direction or pattern. PC2, orthogonal to PC1, contributed additional variability, providing insights into patterns that not accounted for by PC1. PC3, orthogonal to both PC1 and PC2, offers unique information about the remaining variability. The principal components succinctly captured key patterns and structures in the dataset, providing a concise representation while preserving critical information from the original variables.

Performance evaluation of the SVR model in anomaly detection used several metrics, including Accuracy, Precision, Recall, and F1-Score. The anomaly detection performance evaluation of SVR algorithms was carried out on the Normalization Dataset and PCA Dataset. The results of the comparison of the two datasets were further analyzed in the Results and Discussion section.

2.11 Validation of K-fold Cross Validation (K-Fold CV) of SVR RBF Model

Validation was carried out with K-Fold Cross Validation (K-Fold CV) using K = 10 folds on Normalization Dataset and PCA Dataset. With the K-Fold CV the model was trained and tested on each fold of different data to obtain a more reliable estimate of how well the model perform on new data not used in training [18].

3. RESULTS AND DISCUSSION

3.1 Ground Truth Labeling

The ground truth labeling performed on the 100.000 sample data based on Permenkes yielded the following results: 13.650 rows were labeled as anomalies, while 86.350 rows were labeled as normal. Thus, in the entire sample dataset, 13,65% of rows were labeled as anomalies, and 86,35% of rows were labeled as normal.

3.2 Comparison of SVR RBF and MLR Performance Evaluation

The independent and target variables used by the SVR RBF and MLR algorithms are presented in previous Table 1. Comparison of SVR RBF and MLR anomaly detection performance evaluation in the testing phase of Normalization Dataset is summarized in Table 3.

Table 3 Performance Anomaly Detection Evaluation Comparison of SVR RBF and MLR on Normalization Dataset

Testing Phase on Anomaly Definition = Residue > 2 * RMSE					
Algorithm	Model Aliases	Accuracy	Precision	Recall	F1-Score
SVR RBF	SVR N	0,89	0,85	0,23	0,36
MLR	MLR N	0,87	0,96	0,08	0,14

In the training and testing stages, the evaluation of anomaly detection performance showed the fact that the SVR RBF model, namely SVR N, had better anomaly detection performance than the MLR model, namely MLR N. The combination of Accuracy, Precision, Recall, and F1-Score values for SVR N out performed than MLR N.

The SVR RBF algorithm had better performance on anomaly detection than MLR. Therefore SVR RBF hyperparameters and anomaly definitions was tuned to obtain the best combination that results in the most optimal Accuracy, Precision, Recall, and especially F1-Score.

3.3 SVR RBF Hyperparameter Tuning & Anomaly Definition Results on Normalization Dataset

Tuning the RBF kernel's SVR hyperparameters and anomaly definition aimed to enable the model on producing the best combination of Accuracy, Precision, Recall, and F1-Score performance evaluations. The tuning results are shown in Table 4.

Based on the tuning results (Table 4), variable sets number 10, 12, 14, and 16 gave the best results, namely 0,97 for Accuracy, 0,84 for Precision, 0,97 for Recall, and 0,90 for F1-Score. Since four sets of variables provided the best results, the variable set number 10 was used as hyperparameters in detecting anomalies with the SVR algorithm and was used to build SVR models with PCA Dataset.

Table 4 Results of SVR RBF Hyperparameter & Anomaly Definition Tuning

No	C	Epsilon	Gamma	Anomaly Definition = Residue > k*RMSE	Accuracy	Precision	Recall	F1-Score
1	1000	1000	1000	Residue >2*RMSE	0,89	0,86	0,23	0,36
2	1000	1000	1000	Residue > 0,5*RMSE	0,97	0,84	0,96	0,90
3	1000	1000	1	Residue >2*RMSE	0,89	0,86	0,23	0,36
4	1000	1000	1	Residue > 0,5*RMSE	0,97	0,84	0,95	0,89
5	1000	1	1000	Residue >2*RMSE	0,89	0,86	0,23	0,36
6	1000	1	1000	Residue > 0,5*RMSE	0,97	0,84	0,96	0,90
7	1000	1	1	Residue >2*RMSE	0,89	0,86	0,23	0,36
8	1000	1	1	Residue > 0,5*RMSE	0,97	0,84	0,95	0,89
9	1	1000	1000	Residue >2*RMSE	0,89	0,85	0,23	0,36
10	1	1000	1000	Residue > 0,5*RMSE	0,97	0,84	0,97	0,90
11	1	1000	1	Residue >2*RMSE	0,89	0,86	0,23	0,36
12	1	1000	1	Residue > 0,5*RMSE	0,97	0,84	0,97	0,90
13	1	1	1000	Residue >2*RMSE	0,89	0,85	0,23	0,36
14	1	1	1000	Residue > 0,5*RMSE	0,97	0,84	0,97	0,90
15	1	1	1	Residue >2*RMSE	0,89	0,85	0,23	0,36
16	1	1	1	Residue > 0,5*RMSE	0,97	0,84	0,97	0,90

3.4 Comparison of SVR Performance Evaluation of Normalization Dataset and PCA Dataset

Comparison of SVR anomaly detection performance evaluation in Normalization Dataset & PCA Dataset testing phase is shown in Table 5.

Table 5. Comparison of SVR Anomaly Detection Performance Evaluation on Normalization Dataset & PCA Dataset

Testing SVR RBF C=1 Epsilon=1000 Gamma=1000 Anomaly Definition = Residue > 0,5 * RMSE					
Dataset	Accuracy	Precision	Recall	F1-Score	Execution Time (seconds)
Normalization	0,97	0,84	0,97	0,90	18
PCA	0,97	0,84	0,97	0,90	15

Several patterns were found by comparing the evaluation of anomaly detection performance on the Normalization Dataset and the PCA Dataset. In both training and testing the use of RBF kernels with the same parameters (C=1, Epsilon=1000, Gamma=1000) as well as the same anomaly definitions (Anomaly Definitions = Residue > 0,5 * RMSE) resulted in the same Accuracy, Precision, Recall, F1-Score performance. The use of parameter sets and anomaly definitions resulted in Accuracy, Precision, Recall, F1-Score performance close to 1 or optimal.

The PCA Dataset representation had anomaly detection performance equivalent to the Normalization Dataset in terms of Accuracy, Precision, Recall, F1-Score performance. However, the significant difference in execution time, as PCA datasets reduced the data dimensions without sacrificing anomaly detection quality. Thus, the selection between Normalization Dataset and PCA Dataset depends on the computational speed and complexity of the data dimensions.

3.5 Results of K-Fold CV on Normalization Dataset & PCA Dataset

The results of validation of anomaly detection of SVR RBF model with Normalization Dataset and PCA Dataset are shown in Table 6.

Table 6. Results of K-Fold CV on Normalization Dataset & PCA Dataset

K = 10	SVR RBF C=1 Epsilon=1000 Gamma=1000 Anomaly Definition = Residue > 0,5 * RMSE							
	Normalization Dataset				PCA Dataset			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
1	0,97	0,82	0,98	0,89	0,97	0,82	0,98	0,89
2	0,97	0,85	0,95	0,90	0,97	0,85	0,95	0,90
3	0,97	0,87	0,88	0,88	0,97	0,87	0,88	0,88
4	0,97	0,85	0,97	0,91	0,97	0,85	0,97	0,91
5	0,97	0,83	1,00	0,91	0,97	0,83	1,00	0,91
6	0,97	0,83	0,98	0,90	0,97	0,83	0,98	0,90
7	0,97	0,84	0,98	0,90	0,97	0,84	0,98	0,90
8	0,97	0,83	0,97	0,89	0,97	0,83	0,97	0,89
9	0,97	0,82	0,97	0,89	0,97	0,82	0,97	0,89
10	0,97	0,84	0,94	0,89	0,97	0,84	0,94	0,89
Average	0,97	0,84	0,96	0,90	0,97	0,84	0,96	0,90

Validating anomaly detection on Normalization Dataset and PCA Dataset using SVR RBF with parameter listed gave similiar and excellent model performance. Applying K-Fold Cross Validation (K = 10) resulted an average accuracy of 97%, indicating the model's consistently classify normal and anomalous data. Furthermore, anomaly detection metrics such as Precision, Recall, and F1-Score exhibited highly positive results, with average precision around 0,84, recall around 0,96, and F1-Score around 0,90. These values reflected the model's robust capability in identifying anomalies with high precision and sensitivity. The anomaly detection validation showed the fact that the SVR model with RBF kernel on Normalization Dataset and PCA Dataset was sufficiently effective in anomaly identification.

4. CONCLUSIONS

The Anomaly Detection of BPJS Kesehatan Claims by Hospitals with Support Vector Regression algorithm showed the fact that the SVR RBF model had better anomaly detection performance than the MLR model on the same sample data. The results of SVR RBF hyperparameter tuning showed the best combination with values of C=1, Epsilon=1000, Gamma=1000, and Anomaly Definition on Residue > 0,5 * RMSE, resulting in anomaly detection performance with Accuracy 0,97, Precision 0,84, Recall 0,97, and F1-Score 0,90. In addition, SVR RBF model on PCA Dataset and Normalization Dataset had equivalent anomaly detection performance. Apllying the model on PCA Dataset gave faster execution times than Normalization Dataset.

REFERENCES

[1] A. Ratnawati, W. bin Mislan Cokrohadisumarto, and N. Kholis, "Improving the satisfaction and loyalty of BPJS healthcare in Indonesia: a Sharia perspective," *Journal of Islamic Marketing*, vol. 12, no. 7, pp. 1316–1338, 2020, doi: 10.1108/JIMA-01-2020-0005

[2] E. Afrina *et al.*, "Defisit Jaminan Kesehatan Nasional (JKN): Mengapa dan Bagaimana Mengatasinya?," 2020.

[3] R. A. Fattah *et al.*, "Incidence of catastrophic health spending in Indonesia: insights from a Household Panel Study 2018–2019," *Int J Equity Health*, vol. 22, no. 1, Dec. 2023, doi: 10.1186/s12939-023-01980-w

[4] N. Fathurrohman and A. Dewi, "Potential Fraud in The Primary Healthcare," *Jurnal Medicoeticolegal dan Manajemen Rumah Sakit*, vol. 7, no. 3, 2018, doi: 10.18196/jmmr.7373

[5] H. K. Prakosa and N. Rokhman, "Anomaly Detection in Hospital Claims Using K-Means and Linear Regression," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 4, p. 391, Oct. 2021, doi: 10.22146/ijccs.68160

[6] M. Meghana, S. Radhika, and V. S. Kumari, "Anomaly Detection for Vertical Plant Wall System using Novel Support Vector Machine in comparison with Linear Regression for improving accuracy," Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 1–5. doi: 10.1109/iconstem56934.2023.10142459

[7] Q. Ai, S. Liu, L. He, and Z. Xu, "Stein Variational Gradient Descent with Multiple Kernel," Jul. 2021, doi: 10.1007/s12559-022-10069-5. Available: http://arxiv.org/abs/2107.09338

[8] T. Wang, X. Cao, Y. Li, Y. Zhai, and G. Ye, "Real-time anomaly data detection method based on mixed kernel function PSO-SVR," *AIP Adv*, vol. 13, no. 6, Jun. 2023, doi: 10.1063/5.0140105

[9] I. Ivan, T. Roman, G. Michal, Z. Khrystyna, and L. Nataliia, "Input Doubling Method based on SVR with RBF kernel in Clinical Practice: Focus on Small Data." Procedia Computer Science, 2021. doi: 10.1016/J.PROCS.2021.03.075

[10] K. Tscharke, S. Issel, and P. Debus, "Semisupervised Anomaly Detection using Support Vector Regression with Quantum Kernel," Aug. 2023, doi: 10.1109/QCE57702.2023.00075. Available: http://arxiv.org/abs/2308.00583

[11] E. S. Fatahillah Pakpahan, T. Valentine, A. Arixson, and S. A. Batubara, "Analisis Hukum Terhadap Tindakan Pidana Penipuan yang Menyalahgunakan BPJS Kesehatan Berdasarkan KUHP," *Syntax Literate ; Jurnal Ilmiah Indonesia*, vol. 6, no. 10, p. 4967, Oct. 2021, doi: 10.36418/syntax-literate.v6i10.4370

[12] S. Kurniawan, H. S. Disemadi, and A. Purwanti, "Urgensi Pencegahan Tindak Pidana Curang (Fraud) Dalam Klaim Asuransi," *Halu Oleo Law Review*, vol. 4, no. 1, p. 38, Mar. 2020, doi: 10.33561/holrev.v4i1.10863. Available: http://ojs.uho.ac.id/index.php/holrev/article/view/10863

[13] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning*. Cambridge University Press, 2020. doi: 10.1017/9781108679930. Available: https://www.cambridge.org/highereducation/books/mathematics-for-machine-learning/5EE57FD1CFB23E6EB11E130309C7EF98#contents

[14] L. Fahrmeir, T. Kneib, S. Lang, and B. D. Marx, "The Classical Linear Model," in *Regression*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2021, pp. 85–190. doi: 10.1007/978-3-662-63882-8_3. Available: https://link.springer.com/10.1007/978-3-662-63882-8_3

[15] M. Awad and R. Khanna, "Support Vector Regression," in *Efficient Learning Machines*, Berkeley, CA: Apress, 2015, pp. 67–80. doi: 10.1007/978-1-4302-5990-9_4

[16] M. U. Ndubuaku, A. Anjum, and A. Liotta, "Unsupervised Anomaly Thresholding from Reconstruction Errors," 2019, pp. 123–129. doi: 10.1007/978-3-030-34914-1_12

[17] M. A. Mondal and Z. Rehena, "Road Traffic Outlier Detection Technique based on Linear Regression," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 2547–2555. doi: 10.1016/j.procs.2020.04.276

[18] H. B. Moss, D. S. Leslie, and P. Rayson, "Using J-K fold Cross Validation to Reduce Variance When Tuning NLP Models," Jun. 2018, doi: https://doi.org/10.48550/arXiv.1806.07139. Available: http://arxiv.org/abs/1806.07139