Received: December 15, 2023. Revised: February 7, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhancing Prediction Accuracy in an Imbalanced Dataset of Dengue Infection Cases Using a Two-layer Ensemble Outlier Detection and Feature Selection Technique

Amiq Fahmi[1,2] Diana Purwitasari[3,4] Surya Sumpeno[1,3,5] Mauridhi Hery Purnomo[1,3,5]*

[1]*Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[2]*Department of Information System, Universitas Dian Nuswantoro, Semarang, Indonesia*
[3]*University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Indonesia*
[4]*Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Indonesia*
[5]*Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Indonesia*
* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Real-world datasets frequently compromise considerably on noise, resulting in the emergence of outlier data. Detecting and removing outliers in large and imbalanced datasets is a challenging and exciting study in machine learning, especially in healthcare, for accurate prediction. Therefore, it is essential to handle outliers properly, as their presence in classification datasets leads to more difficult, inaccurate, and lower predictive modelling performance. The study proposes methods to enhance prediction accuracy in an imbalanced real-world health dataset of dengue infection cases. First, use a two-layer ensemble method called IsFLOF, which involves an isolation forest (IsF) and a local outlier factor (LOF) to find and accurately eliminate global and local outliers. This approach overcomes the limitations of the IsF algorithm, which is only sensitive to global outliers but vulnerable to local outliers, while LOF excels in local outlier detection but has high complexity. Second, once a dataset with correctly measured value distributions was obtained by eliminating outliers, a resampling process was conducted to prevent prediction bias caused by imbalanced instance data in the multi-class setting. Subsequently, insignificant features were filtered out to further refine the dataset. In the end, eight machine learning algorithms are used to test the robustness and effectiveness of the proposed method. The experimental results showed that the AdaBoost classifier, combined with selected features from the Fast Correlation-Based Filter (FCBF), achieved 93.5% and 95.1% accuracy in training and testing, respectively. In a more distant context, the proposed method is tested and compared with recent methods, including using a public dataset of imbalanced hypothyroid cases. It showed higher and more acceptable prediction accuracy than the original and synthetic data.

Keywords: Classification accuracy, Outlier detection, Imbalanced dataset, Resampling, Feature selection, Dengue infection cases.

1. Introduction

Data mining is an active area of research that experimenters are increasingly utilizing to analyze expansive medical and public health datasets and develop robust prediction systems [1]. However, in many real-world cases where a machine learning model is relied upon to handle large data sets, such as supporting the clinical diagnosis of arboviral diseases [2], especially dengue infection cases [3, 4], there are still several issues that need to be further explored and handled. Among the challenges is that real-world datasets are troubled by noise, exaggeration, and imbalance [4, 5]. This situation affects the emergence of outliers, distinct behaviors from the majority within the same class with exceptional values, skew classes, and insignificant features, leading to distortions of data patterns and trends and prediction bias [5, 6]. As a result, the prediction task becomes more complicated, affecting the model's performance and accuracy [7]. For this reason, identifying and

removing outliers is a very challenging task in practice to produce classification datasets with optimal performance and accuracy.

In most literature publications, constructing robust classification models involves improving data quality by mining excess outliers, resolving imbalanced instances of class data, and filtering out irrelevant features. The purpose is to enhance model training efficiency and prediction accuracy [4, 8, 9]. Due to the large dataset with numerous features, an outlier detection algorithm is required to minimize errors in checking and correcting the data. Several outlier detection algorithms are renowned for their potential for mining outliers in large data sets, predominantly isolation forest (IsF), local outlier factors (LOF), and linear models of PCA, which are One-class SVM [10]. In comparison, IsF uses a decision tree technique to separate outliers from a dataset until the outliers are entirely isolated. Although IsF is swift and efficient in detecting global outliers, it is soft against local outliers and has the potential for false-positive detection. Instead, LOF is density-based, highly complex in processing and time, but efficiently identifies local outliers [11]. Meanwhile, one-class SVM performs better with a non-Gaussian distribution than a Gaussian. At the same time, machine learning has attracted interest in combinatorial investigations for the most effective prediction models on specific data sets [12]. Thus, there is a wide-open possibility of using an ensemble approach to handle outliers by improving the performance of these techniques before applying an appropriate model classification [11, 13]. The ensemble method, like kNN-LOF [12], offers a solidity for sequential model construction by correcting the falsehoods of previous models to enhance performance and accuracy.

Also, to address imbalanced class instance data, resampling techniques have been recognized for their significant impact on tackling prediction bias [4], specifically random resampling, random under- and over-sampling [14, 15], and SMOTE [16] techniques. Expressly, data levels are a crucial technique for resolving imbalanced data sets. Better thoughtfully, several feature reduction techniques are prioritized, including filter-based, wrapping, and embedding approaches to further increase efficiency and accuracy by pruning dataset complexity from insignificant attributes [9, 17, 18]. Thus, the negative impact of the statistical analysis can be avoided, and optimal and reliable results can be achieved [7, 19].

Dengue infection refers to a group of diseases caused by the dengue virus in humans. Prediction accuracy is crucial to supporting medical professionals in detecting, diagnosing, and treating dengue infections appropriately to avert the grave risk of patient fatality. A study that has been conducted by Fahmi et al. [3] utilized a real-world dataset of dengue infection patients to challenge the accuracy of prediction results for dengue fever, dengue hemorrhagic fever, and dengue shock syndrome using the ReliefF filter. However, the accuracy result is low, at 72.4%. The data set implies the presence of outliers and an imbalance of data class instances with skewed distributions of 34.7%, 61.0%, and 4.3% in each class, respectively. Hence, it is indisputable that controlling the impact of outliers, balancing data, and filtering attributes remains an essential challenge in machine learning for enhancing classification performance and accuracy in areas such as diabetes prediction [20] and dengue infection [3, 4].

To solve the problem of outliers in an imbalanced real-world dataset and enhance the prediction accuracy of dengue infection case classification optimally, the contribution of this paper is as follows:

(1) This study proposed a two-layer ensemble method, called IsFLOF, that involves an isolation forest (IsF) and local outlier factors (LOF) to handle and eliminate the global and local outliers effectively and appropriately. The LOF enhances IsF precision in outlier mining, reducing complexity, finalizing the training dataset, and revamping new sample observations.
(2) Random resampling techniques are utilized to address imbalanced datasets.
(3) Filter-based feature selection techniques are enforced to refine the dataset further.

The three methods are proposed as a unified set, a recent approach to improving classification accuracy in the case of imbalanced datasets with enormous noise and emerging outliers.

In the final investigation, we use eight machine learning algorithms to test the proposed method's robustness, efficiency, and effectiveness. These are Naïve bayes, decision tree, K-nearest neighbor, random forest, neural network, AdaBoost, support vector machine, and logistic regression. In a set of experiments, we used the original dataset of dengue infection cases, the public dataset of hypothyroid cases, and the recently available technique kNN-LOF [12] for comparison. Effectiveness is measured based on standard metrics such as accuracy, precision, recall, F-score, and AUC. Eventually, the proposed method, IsFLOF, delivers more accurate results than outlier detection techniques using IsF and LOF individually and kNN-LOF on primary and synthetic datasets.

The structure of this paper is as follows: section 2 provides a comprehensive review and explanation of

related works, including outlier detection techniques, resampling, feature selection, and classification problems. Section 3 presents the methodology proposed, while section 4 presents the experimental results and provides a space for discussion. Lastly, section 5 explicitly highlights conclusions and future work.

2. Related work

Outlier detection techniques have provided valuable insights into data mining, especially in the preprocessing stage. This is due to the failure of basic statistical approaches to handle most machine-learning datasets that contain numerous features [12]. As an effect, several outlier detection algorithms are used as pipelines in modelling, mainly to improve the performance and robustness of the model. Several outlier detection techniques, such as isolation forest, LOF, One-class SVM, DBSCAN, K-Means, and others, have been widely used in mining outliers in large datasets in several cases, similarly preventing credit card fraud and data leakage [21], tumor classification, breast cancer detection, patient monitoring through ECG signals [22, 23], and studying metabolism [24]. The authors used historical datasets from different sources, such as Kaggle, the ML repository at UCI, and real-world datasets, to analyze and create synthetic datasets for training models. The results indicate that mining outliers optimally and removing them can improve prediction accuracy significantly compared to standard classification methods when detecting credit card fraud and other anomalies [20-23].

Selecting an appropriate outlier detection technique sensitive to noise and outliers is challenging because each has drawbacks, such as the K-Means algorithm, which is sensitive to cluster center initialization and ineffective for data with complex cluster shapes or uneven distributions [25]. One-Class SVM is sensitive to parameter selection and large data size [26]. DBSCAN, a method for spatial clustering applications with noise, has drawbacks like sensitivity to parameters, ineffectiveness on diverse data densities, and vulnerability to outliers, especially low cluster density being considered noise or outliers [27].

The local outlier factor (LOF) is one of the most widely used density-based methods for automatically mining outliers. Sugidamayatno and Lelono [28] conducted an outlier detection analysis on transaction data from credit card customers using the INFLO, AFV, and LOF algorithms. As a result, the LOF algorithm has a higher accuracy value, exceeding INFLO and AFV. However, LOF is unsuitable for large-scale datasets due to its high processing and time complexity. In comparison, there is the Isolation Forest (IsF) method. IsF uses a decision tree technique to separate outliers from a dataset until the outliers are entirely isolated. Gao et al. [29] reasoned that the IsF technique outperforms LOF algorithms due to its low time complexity and better anomaly quantity. However, the algorithm's accuracy is reduced due to its weakness in identifying local outlier points and its inability to avoid false positives. Therefore, the IsF algorithm was suggested to improve further in identifying both outlier points. Contrastingly, ensemble methods were developed to construct models where each model improves the other. Alsini et al. [13] proposed the Isolation Forest technique based on the sliding window for the local outlier factor, which efficiently detects those outliers at the concrete mix design stage. Xu et al. [12] introduced kNN-LOF, a new outlier detection algorithm, to improve the accuracy of existing methods and overcome their limitations. In addition, Cheng et al. [11] utilized IsF, LOF, and their combination techniques on several datasets. The combination of IsF and LOF outperformed IsF and LOF individually on synthetic datasets with more than 98% accuracy and 72% on real-world datasets. However, selecting and applying a particular ensemble method should consider complexity, sensitivity, and interpretability to outliers [12,30]. Therefore, there is an openness to proposing a new ensemble approach and comparing it with existing ones.

Also, to further improve the efficiency of the training model and accuracy, Cherrington et al. [17] thoroughly analyzed filter-based feature reduction techniques using ranking procedures, specifically focusing on information gain, Chi-square, and ReliefF. The point of concern is determining ranking-based thresholds, particularly in the big-data era, where filtering uses a limited number of attributes. In line with Yusuf et al. [23], they combine feature selection with outlier detection techniques to enhance breast cancer diagnosis accuracy. They used seven machine-learning methods on the Wisconsin dataset. When outliers are removed from the dataset and attributes are filtered out, the test results show that Random Forest, AdaBoost, and Logistic Regression classifiers are 99.12% accurate. In addition, Thabtah et al. [31] used the feature selection technique to evaluate 27 data sets. They assessed the performance of information gain and Chi-square techniques suitable for large data sets. In the future, they introduced the fast correlation-based filter technique, which efficiently identifies features and redundancies in big data without pairwise correlation analysis.

For more arguments, the study by Fahmi et al. [3] focused on classifying dengue fever using the real-world dataset of dengue infection cases paired with the ReliefF feature selection technique. They presented their findings with a low accuracy of 72.4%. Correspondingly, they then used weighting methods, resampling techniques, and feature selection to improve the accuracy further. They tested its accuracy using eight algorithms, including NB, DT, KNN, random forest, NN, AdaBoost, SVM, and logistic regression. The results showed a significant improvement in accuracy of 87% [4]. Also, Mello-Román et al. [32] classified dengue into binary categories as "severe" and "not severe" by using NB, KNN, Rule Bayes, ID3, and DT classification algorithms. The NB algorithm produces a maximum accuracy of 72%. As for comparison, Guleria et al. [33] used algorithms such as DT, Random Forest, NB, and deep learning ANN to predict early hypothyroidism with multi-class classification target features. The DT and Random Forest performed better, with the highest accuracy of 99.56% and 99.31%, respectively. On the other hand, Chaganti et al. [34] presented the prediction of hypothyroid cases using a Random Forest-based feature approach, and the highest accuracy achieved was 99% for ten thyroid diseases.

3. Proposed methodology

Based on related research, in this study, we propose three influential methods to improve prediction accuracy on an imbalanced primary dataset of real-world dengue infection cases.

Primarily, we propose a two-layer ensemble method called IsFLOF. This method combines the strengths of the isolation forest (IsF) and local outlier factor (LOF) algorithms to handle and eliminate outliers accurately with low time complexity. Models are constructed sequentially, where the second model improves on the weaknesses of the first model to produce more accurate and stable predictions. The IsF algorithm is used in the first layer to swiftly scan the dataset and isolate outliers to produce a set of candidate outliers. The LOF algorithm effectively filters out the candidate dataset's outliers in the second layer based on the outlier coefficient and pruning threshold values. The LOF method ensures high-quality results without outliers and detects new data samples as a final validation step, improving prediction accuracy. Secondarily, overcome the problem of imbalanced datasets in a multi-class setting using random resampling techniques. In the third position, we picked out the essential features

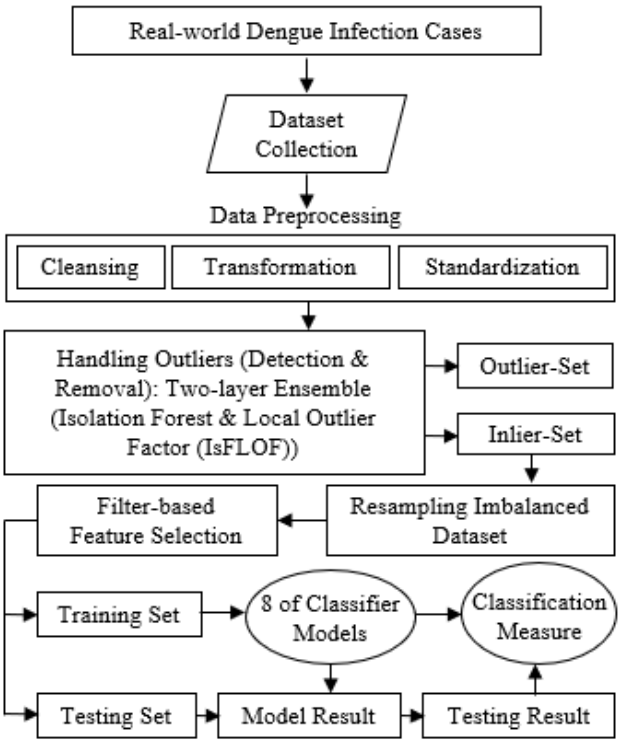


Figure. 1 Proposed method

using filter-based techniques, which are compatible with any machine learning model, especially the information gain, Chi-square, ReliefF, and fast correlation-based filter.

The system design proposed is sequentially shown in Fig. 1. In this study, the concept of "inlier" refers to data values that fall within the measured part of the distribution. Meanwhile, "outlier" denoted data points that deviated significantly from the distribution.

The stages of the study began with collecting real-world datasets, preprocessing data, handling outliers, resampling imbalanced datasets, selecting significant features, partitioning the dataset into training and test sets, and classifying them using eight machine learning algorithms. Experiments were conducted on the original dataset of dengue infection cases and a public dataset of hypothyroid cases as a comparison. Effectiveness is measured based on standard metrics such as accuracy, precision, recall, F-score, and AUC. In conclusion, the accuracy results of the proposed method are compared with outlier detection methods like kNN-LOF, IsF, and LOF on original and synthetic datasets.

3.1 Data collection

Real-world data on dengue infection cases was collected from diverse hospitals and 37 community health centers in 16 subdistricts of Semarang City. The data for each patient has been investigated and

verified by epidemiologists and health professionals at semarang city health office in central java, Indonesia, from 2016 to 2019 on vector-transmitted diseases and zoonoses. The dengue dataset was validated based on three clinical diagnosis criteria, specifically DF (dengue fever), DHF (dengue hemorrhagic fever), and DSS (dengue shock syndrome). The dataset had 1 output attribute for the clinical diagnostic criteria and 16 distinct input attributes. The dataset contained 9 categorical attribute types, including 1) Sex (male/female), 2) R/L test (Positive/Negative), 3) Pleural effusion (yes/no), 4) Ascites (yes/no), 5) Hypoproteinemia (yes/no), 6) Hepatomegaly (yes/no), 7) Shock (yes/no), 8) IgM (positive/negative), and 9) IgG (positive/negative). Also, the remaining 7 numeric attributes were 1) Age (> 0), 2) Period of symptoms (0-14), 3) Period of diagnosis (0-7), 4) Thrombocytes (1000-600000), 5) Initial hematocrit (11-70), 6) Diagnosis of hematocrit (11-70), and 7) Hemoglobin (4.5-25.4). Investigating the measures of hematocrit and thrombocyte quantity is an indicator for diagnosing dengue infection cases. The hematocrit value usually increases (hemoconcentration). Otherwise, the thrombocyte quantity will decrease (thrombocytopenia), indicating the severity stage. The dataset had 14,044 data samples, which showed an uneven distribution of classes with a ratio of 4,875 (4.7%): 8,560 (61.0%) and 609 (4.3%) for the DF, DHF, and DSS classes, respectively.

3.2 Data preprocessing

Data preprocessing concerns the process of preparing data for training and testing. The initial preprocessing stages encompassed data cleansing, transformation, and standardization.

Data cleaning is a process to ensure the completeness and sensibility of datasets used in training and testing. Statistical techniques like mean values fill in missing or unknown data values. At the same time, non-statistical imputation is applied to attributes like blank hematocrit values, calculated using a formula three times the hemoglobin level [35]. Another method involves calculating the distribution of values and randomly selecting them.

Data transformation involves converting data from one form to another to improve processing efficiency by changing positive/negative or yes/no values to 1 and 0, such as in attributes "R/L Test" and "Pleural effusion".

Data standardization maintains consistency among datasets, bringing variations to a standard scale where the average deviation was 0 and the standard deviation became 1 $\left(\mu = 0, \sigma^2 = 1\right)$. This process facilitated the interpretation of disparities across different data sources or variables, such as Age, Period of symptoms, Period of diagnosis, initial hematocrit, Diagnosis of hematocrit, and Hemoglobin.

Algorithm 1: Isolation forest (IsF)

Input : X-input dataset, t-number of trees, s-sub sampling size
Output : a set of t iTrees
Step:
1: Initialize Forest
2: set height limit $l = ceilinglog_2 s$
3: for $i = 1\ to\ t$ do
4: $X' \leftarrow sampleX, s$
5: $Forest \leftarrow$ Forest $\cup$ iTree$X', 0, t$
6: end for
7: return $Forest$

Figure. 2 The algorithm of IsF technique

3.3 Handling outliers: Uses IsF, LOF, the two-layer ensemble of IsFLOF, and kNN-LOF techniques.

The isolation forest (IsF) property was used to find candidates for outliers in the dataset. This technique was done by calculating the density distance between data instances. The algorithm used binary decision trees randomly constructed from a collection of data instances to explore each tree and compute outlier scores for each data instance point. The isolation tree algorithm construction, represented by the (X, e, h) f function, was defined with X as the input dataset, e as the current tree height, and h as the height limit [11]. The IsF construction algorithm was implemented in stages described by Algorithm 1 in Fig. 2.

On the other hand, the LOF algorithm was used to detect data density-based unsupervised outliers by calculating the local deviation score of specific data points. Outliers were found by looking at how solid the connections were between each data point and its neighboring points. When the point density was reduced, the algorithm became more identifiable as an outlier [11,36]. The more degraded the dot compactness, the more likely it was to be admitted as an outlier. LOF algorithm settings are based on definitions 1–6. The steps are described in Algorithm 2 in Fig. 3.

Definition 1: (dp, q): distance between points p and q.
Definition 2: $k - distance$: rank the distance from point p to other data points, and the distance from point p to data point k is recorded as $k - distp$.

Definition 3: k nearest-neighbors: data point is set to point p distance less than $k - distp$, recorded as $N_k p$.

Definition 4: reach distance using Eq. (1):

$$reach - dist_k p, r = max\{k - dist_r, dp, r\} \quad (1)$$

Definition 5: local reachability $densitylrd$: The reciprocal of the average reachable distance of data points p and k of its nearest neighbors, which is calculated by Eq. (2).

$$lrdp = 1 \frac{s \in N_k p^{reach - dist_k p, r}}{|N_k p|} \quad (2)$$

Definition 6: LOF: the mean of the ratio of the local get at able compactness of the point p neighborhood point to the point's attainable density p, calculate using Eq. (3).

$$lofp = \frac{t \in N_k p \frac{lrdt}{lrdp}}{|N_k p|} \quad (3)$$

The IsFLOF two-layer ensemble technique employs the IsF algorithm to mine outlier candidates from the original data set for further processing using the LOF algorithm [7, 11, 19]. It is pruning several normal data points according to the threshold (<0.5) based on the definition of the outlier coefficient (0–1) by calculating the dataset $D = \{d1, d2, \ldots, dn\}$. Where n is the number of samples from D. D_i is an attribute in D, and $d_i = \{x_1, x_2, \ldots, x_n\}$. x_j is the specified data value of the d_i attribute. The attribute outlier coefficient is defined in Eq. (4).

$$fd_i = \frac{\sqrt{\frac{x_j - \bar{x}^2}{n}}}{\bar{x}} = \frac{\sqrt{\frac{x_j - \bar{x}^2}{n}}}{n\bar{x}^2} \quad (4)$$

$\bar{X}$ is the mean of the d_i attribute, and fd_i is used to measure the dispersion of the d_i attribute. The outlier coefficient is calculated for each attribute in the dataset. To get the outlier coefficient D_f attribute vector from the dataset, use Eq. (5).

$$D_f = fd_1, \ldots, fd_n \quad (5)$$

Based on the outlier coefficient vector, the amount of inaccuracy can be calculated as threshold θ_D, representing the proportion of outliers in the dataset. Eq. (6) shows that Top_m refers to the value of m with the magnitude of the dispersion coefficient, and α is the adjustment factor. Both m and α depend on careful consideration of the size and distribution of the data set.

Algorithm 2: Local outlier factor (LOF)

Input : k-number of near neighbors, m-number of outliers, D-outlier candidate dataset.
Output: $topm$ outliers.
Step:
1: for j= 1 to $lenD$ do
2: compute $k - dstp$
3: compute $N_k p$
4: end for
5: calculate $reach - dist_k p, r \; and \; lrdp$
6: calculate $lofp$
7: sort the lof values of all points in descending order
8: return the m data objects with the large lof values, which are the outliers

Figure. 3 The algorithm of LOF technique

$$\theta_D = \frac{\alpha Top\_mD_f}{m} \quad (6)$$

The outlier score of each point calculated by IsF, $1 - \theta_D$ data points from the data set is pruned, with the remaining data points constituting the candidate outlier sets.

kNN-LOF is a density-based outlier detection method that improves distance-based global outlier detection. It assigns local outlier coefficients to data objects, indicating their degree of outliers relative to their neighborhood. The LOF calculation method evaluates the ratio of local density to average density, with a LOF value close to 1 indicating even density. The larger the LOF value, the more likely the object is an outlier [12].

Furthermore, these IsF, LOF, IsFLOF, and kNN-LOF techniques were used to identify outliers in real-world and public datasets imbalanced in dengue infection [3,4] and hypothyroid cases [33,34]. The dataset of primary dengue infection cases consists of 16 input attributes and 1 output attribute. It contains 14,044 data instances and exhibits imbalanced distributions of 4,875, 8,560, and 609 for the output attributes representing dengue fever (DF), dengue hemorrhagic fever (DHF), and dengue shock syndrome (DSS) classes, respectively. The hypothyroid dataset consists of 29 input attributes, including age, sex, on thyroxine, query on thyroxine, on antithyroid medication, sick, pregnant, thyroid surgery, I131 treatment, query hypothyroid, query hyperthyroid, lithium, goitre, tumor, hypopituitary, psych, TSH measured, TSH, T3 measured, T3, TT4 measured, TT4, T4U measured, T4U, FTI measured, FTI, TBG measured, TBG, referral source. There are a total of 3,772 data instances in the dataset, with an

Table 1. Outlier detection uses IsF, LOF, IsFLOF, and kNN-LOF in the original dataset of dengue infection cases

Outlier Detection Technique	Result					
	Inlier			Outlier		
	DF	DHF	DSS	DF	DHF	DSS
IsF	4,485	7,834	557 12,876	390	726	52 1,168
LOF	4,558	7,759	322 12,639	317	801	287 1,405
Two-layer Ensamble IsFLOF	4,461	7,822	531 12,814	414	738	78 1,230
kNN-LOF	4,488	7,826	514 12,828	387	734	95 1,261

Table 2. Outlier detection uses IsF, LOF, IsFLOF, and kNN-LOF in the original dataset of hypothyroid

Outlier Detection Technique	Result							
	Inlier				Outlier			
	CH	N	PH	SH	CH	N	PH	SH
IsF	186	3,307	89	1 3,583	8	174	6	1 189
LOF	179	3,273	92	1 3,545	15	208	3	1 227
Two-layer Ensamble IsFLOF	178	3,275	93	2 3,548	16	206	2	0 224
kNN-LOF	191	3,179	83	2 3,455	3	302	12	0 224

imbalance in each class. The negative class (N) has 3,481 instances, while the compensated hypothyroid (CH), primary hypothyroid (PH), and secondary hypothyroid (SH) classes have 194, 95, and 2 instances, respectively. The mining outcomes for both datasets are displayed in Tables 1 and 2.

3.4 Resampling

The random resampling technique solves imbalanced class instances in a dataset by combining over- and under-sampling strategies. This study adopted the unsupervised filter resample instance technique from Weka 3.8.5, an open-source program [37], which is applied to data-level solutions.

3.5 Feature selection

This research explores filter-based feature reduction techniques, especially information gain, Chi-square, ReliefF, and FCBF, compatible with any machine learning model, to improve model efficiency and accuracy and prevent overfitting [17]. The information gain (IG) technique measures feature relevance and influence, potentially reducing dimensions by assessing entropy reduction before and after separation [38, 39]. By calculating a feature's entropy, the best attribute is identified.

Chi-square is a statistical technique that evaluates the dependency of a feature on class value. Suppose the occurrence part did not depend on the class value, discarded [40]. In contrast, the feature was considered significant. The chi-square value was calculated using metrics such as true positive (TP), false positive (FP), true negative (TN), false negative (FN), and the probability of the number of positive cases, as well as the likelihood of the number of negative instances.

ReliefF is an algorithmic approach that assigns weights to each feature based on its correlation with the data instance class. The algorithm evaluates distinctions in feature weights and removes features that fall below a certain threshold [39].

The fast correlation-based filter (FCBF) is a multivariate feature selection technique that measures interactions between features and selects the best subset based on correlation coefficient denomination. The FCBF algorithm selects high-correlation features with class, focusing on predictive data instance class features, using Symmetrical Uncertainty (SU) to calculate overall feature scores [41]. SU corrects bias and normalizes the Information Gain (IG) algorithm in feature selection. SU makes feature relevance comparisons fairer.

3.6 Data splitting

The experimental data set is divided into a training and testing set by 70% and 30%, respectively. Data splitting was meant to prevent overfitting. The $k-fold=10$ technique is recommended for model validation, dividing the movement set into 10 equal-sized folds [3,4].

3.7 Classification model

In this section, we provide a brief overview of the eight algorithms used to test our proposed method, specifically Naïve Bayes (NB), decision tree (DT), K-nearest neighbor (KNN), random forest, neural network (NN), AdaBoost, support vector machine (SVM), and logistic regression.

KNN is a classification algorithm that predicts by determining the k nearest neighbors from n training samples. The algorithm predicted the average and decided how many nearby data instances to look at in the feature space. The KNN technique was selected due to its resistance to noisy data. The parameter settings of the algorithm were k=5, metric=euclidian, and weight=uniform [42].

NB is a simple, efficient, and fast probabilistic classifier algorithm that combines Bayes' theorem with the feature independence assumption. The NB algorithm calculated the probability value of each class based on a set of features with the highest probability. Moreover, the algorithm was selected because it achieved maximum precision with minimal training data and was effective with high-dimensional datasets [43].

The SVM classifier used a hyperplane to divide the attribute space and maximize the margin between various class occurrences. It mapped inputs to a higher-dimensional feature space, producing the highest predictive performance. Setting the parameters cost (C), regression loss epsilon (ε), and an appropriate kernel, such as linear, polynomial, RBF, or sigmoid, were necessary for the estimation to be accurate [44].

The DT algorithm [45] is performed by presuming pruning. It was used for classification and regression assignments on numerical and categorical datasets. DT recursively divided the data into subsets based on the most significant features, creating internal nodes representing features and leaf nodes representing predictions based on class purity (information gathering for categories and mean squared error for numerical target variables). Each internal node represented a test on a feature, and each branch indicated a test result. Default parameters were used, such as the minimum number in leave=2, not splitting subsets smaller than 5, limiting the maximum tree depth to 100, and stopping the majority when it reached 95%.

Logistic regression is a classification algorithm incorporating Lasso (L1) or Ridge (L2) regularization. It was explicitly designed for classification tasks in data analysis. The choice of parameters depended on C . Where $C > 1$ represented a strong parameter and $C < 1$ was weak. However, using the balanced class distribution weighting option could decrease performance [44].

Random Forest is a highly effective classifier that efficiently handles large datasets [46]. It predicted outcomes using an ensemble of decision trees and non-parametric patterns to simplify the complexity of the probability density. Random forest constructed a set of decision trees, each considering a subset of attributes. The best feature was randomly selected based on the majority vote of the independently developed trees in the forest. The fundamental parameters included the $number\ of\ trees = 10$, the number of attributes considered at each $split = 5$, the maximum depth of individual $trees = 3$, and the avoidance of splitting subsets smaller than 5.

Table 3. Classification performance measurement

Measure	Formula
Accuracy	$(TP + TN)/All$
Precision	$TP\ /\ (TP\ +\ FP)$
Recall (Sensitivity, TPR)	$TP\ /\ (TP\ +\ FN)$
Specificity (TNR)	$TN\ /\ (TN\ +\ FP)$
Balanced Accuracy	$(Sensitivity\ +\ Specificity)\ /2$
F1 Score	$2\ x\ (Precision\ x\ Recall)\ /\ (Precision\ +\ Recall)$

The NN was developed as a multilayer perceptron (MLP) classifier with a backpropagation algorithm to learn non-linear and linear models. The optimal performance of NN algorithm was achieved by approximately setting model parameters, such as the number of neurons in the hidden layer (100), the activation function for the hidden layer (Relu, logistic, tanh, identity), the weight optimization solver (Adam, SGD, L-BFGS-B), the alpha-L2 penalty (regularization term) with $\alpha = 0.0001$, and the maximum number of iterations = 200 [46].

AdaBoost was developed as an algorithm for addressing both classification and regression problems. It combined multiple "weak" classifiers to create a "strong" one by assigning higher weights to misclassified data points and training the next classifier using these weighted data points. The final prediction was made through a weighted majority vote of all individual classifiers. AdaBoost was considered an "Adaptive" algorithm due to its dynamic adjustment of weights [47].

3.8 Accuracy matrix

Accuracy metrics were used to measure the performance of all techniques proposed, both during the training and testing phases of the classifier model [3,4], as shown in Table 3. The confusion matrix provided reference information for four criteria, specifically accuracy, precision, recall, and F-measure, effectively demonstrating the ground truth. It visualized the comparison between actual and predicted values.

Accuracy is a metric that describes the proportion of all correct predictions across all classes. Precision measures the model's accuracy in classifying a sample as positive. Recall measures the percentage of a model's ability to detect positive samples. A weighted harmonic average model that combines precision and recall is evaluated using F-measure to determine its accuracy. At the same time, the Area under the curve (AUC) of the receiver operating characteristic curve (ROC) determines a model's specificity and sensitivity, with a value closer to 1 indicating better data fit.

4. Experiment result and discussion

This section presents and discusses the experimental results using the dengue infection [3,4] and hypothyroid case [33, 34] datasets, both in the original dataset and those mined from outliers and resampled. The classification process employs eight classifier algorithms, specifically NB, DT, KNN, Random Forest, NN, AdaBoost, SVM, and Logistic Regression.

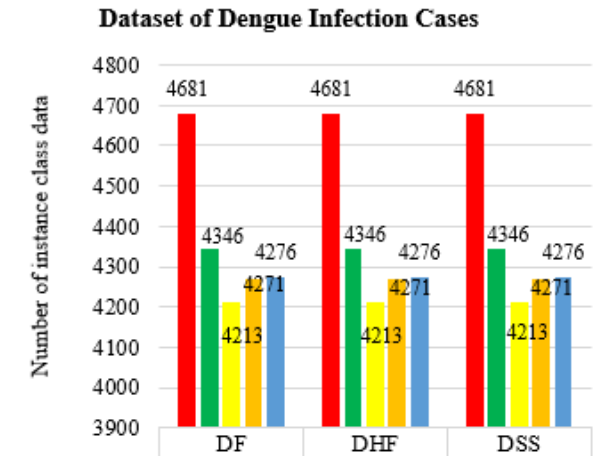
The experiment implicates various scenarios. Firstly, we used the original dataset and the dataset mined from outliers using IsF, LOF, IsFLOF, and kNN-LOF techniques for comparison, as in Tables 1 and 2. Secondly, we use a dataset that has been balanced using the random resampling technique, as visualized in Figs. 4 and 5. The strategy also involves filter-based substantial feature selection techniques using information gain, Chi-square, ReliefF, and FCBF in each training scenario.

Before the classification process, the datasets were divided into 70% and 30% for training and testing sets, respectively. The k-fold cross-validation technique with a value of $k = 10$ was employed.

The performance results of the eight classifier models, obtained after training and testing using the various scenarios, are presented in table form. The table included accuracy (CA), precision (Prec.), recall (Rec.), F1-measure (F1), and area under the curve (AUC) values. However, it only showed the highest accuracy results among the dataset's three significant classifiers involving filter-based feature selection. The performance of all classifier algorithms was compared by examining the test score results and referencing the ground truth information from the confusion matrix. A comprehensive interpretation of the results, in conjunction with several previously published papers, allowed for a thorough analysis at this stage.

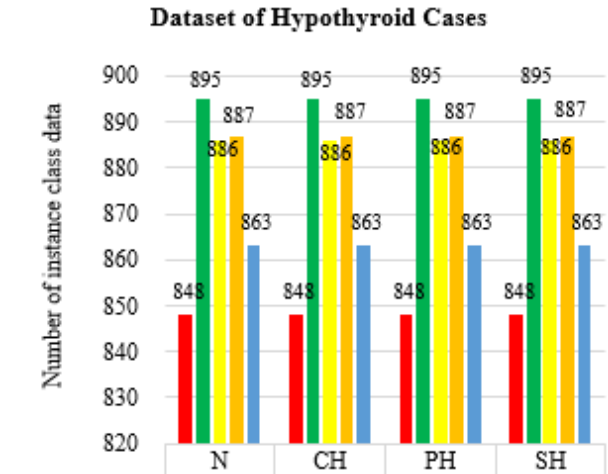
Table 4 displays the order of prediction accuracy results from the three classifiers in the original dataset of dengue infection cases. The NN classifier with ReliefF feature selection achieved the highest accuracy of 72.0% and 72.4% in training and testing. Meanwhile, the logistic regression algorithm and Random Forest classifier showed stability and superiority in the information gain feature selection technique. However, the accuracy of those results still needs to be enhanced, with an AUC of more than 80%.

In the subsequent experiment, we used the original dataset of dengue infection cases mined using IsF, LOF, IsFLOF, and kNN-LOF outlier detection methods. Tables 5-8 show the results. Table



	DF	DHF	DSS
Primary Data	4681	4681	4681
IsF	4346	4346	4346
LOF	4213	4213	4213
IsFLOF	4271	4271	4271
kNN-LOF	4276	4276	4276

Figure. 4 Dataset of dengue infection cases with outlier detection and resampling



	N	CH	PH	SH
Primary Data	848	848	848	848
IsF	895	895	895	895
LOF	886	886	886	886
IsFLOF	887	887	887	887
kNN-LOF	863	863	863	863

Figure. 5 Dataset of hypothyroid cases with outlier detection and resampling

5 shows that the NN classifier has superior accuracy with 83.04% and 84.12% in training and testing by utilizing Chi-Square feature reduction on the dataset mined using the IsF technique. The logistic Regression classifier achieved the second highest accuracy using the FCBF feature reduction technique, while random forest performed best with ReliefF feature reduction.

Further classification results using the LOF outlier detection technique, and the results are presented in

Table 4. High accuracy results of 3 significant classifiers in the primary dataset of dengue infection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Neural Network	ReliefF	0.784	0.720	0.712	0.714	0.719	0.789	0.724	0.718	0.722	0.726
Logistic Regression	Information Gain	0.779	0.712	0.695	0.711	0.712	0.785	0.717	0.699	0.717	0.717
Random Forest	Information Gain	0.767	0.713	0.711	0.713	0.713	0.771	0.716	0.713	0.715	0.716

Table 5. High accuracy results of 3 significant classifications in the primary dataset of dengue infection using IsF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Neural Network	Chi Square	0.879	0.830	0.819	0.817	0.830	0.856	0.841	0.830	0.829	0.841
Logistic Regression	FCBF	0.874	0.829	0.817	0.816	0.829	0.852	0.837	0.825	0.824	0.837
Random Forest	ReliefF	0.874	0.829	0.817	0.816	0.829	0.852	0.837	0.825	0.824	0.837

Table 6. High accuracy results of 3 significant classifications in the primary dataset of dengue infection using LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Neural Network	Chi Square	0.891	0.824	0.819	0.820	0.824	0.899	0.834	0.827	0.830	0.834
Logistic Regression	ReliefF	0.888	0.823	0.809	0.821	0.823	0.893	0.838	0.826	0.837	0.838
Random Forest	ReliefF	0.863	0.815	0.812	0.813	0.815	0.883	0.829	0.827	0.828	0.829

Table 7. High accuracy results of 3 significant classifications in the primary dataset of dengue infection using IsFLOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Neural Network	Chi Square	0.931	0.870	0.863	0.866	0.870	0.944	0.883	0.875	0.880	0.883
Logistic Regression	Information Gain	0.926	0.864	0.846	0.863	0.864	0.925	0.874	0.871	0.872	0.874
Random Forest	FCBF	0.915	0.863	0.859	0.862	0.863	0.919	0.872	0.869	0.870	0.872

Table 8. High accuracy results of 3 significant classifications in the primary dataset of dengue infection using kNN-LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Logistic Regression	Information Gain	0.872	0.775	0.771	0.772	0.775	0.873	0.778	0.772	0.776	0.778
	ReliefF	0.870	0.775	0.771	0.772	0.775	0.871	0.778	0.772	0.776	0.778
	FCBF	0.872	0.774	0.770	0.772	0.774	0.873	0.778	0.772	0.776	0.778

Table 6. The test results show that the NN classifier has better accuracy results of 82.43% on the training set than other classifiers and 83.46% on the testing set when the Chi-Square feature reduction technique is used. Even if the accuracy value during testing was 0.4% lower than the ReliefF feature reduction technique, it is worth noting that the results showed a meaningful increase of 1% over the training accuracy. Compared to the IsF mining technique, the results show that LOF has low accuracy.

This study introduces a refreshed approach, a two-layer ensemble outlier detection technique called IsFLOF, which is specialized to mine outliers precisely, effectively, and efficiently. In addition, we also present an existing and recent outlier detection technique, kNN-LOF, for comparison. Tables 7 and 8 display the test results for dengue infection cases using IsFLOF and kNN-LOF outlier detection techniques. The IsFLOF outlier mining technique in Table 7 demonstrates a classification accuracy of 87,04%, much higher than the original dataset and the dataset mined using the IsF, LOF, and kNN-LOF strategies. The increase in average accuracy is 7.6%, and the AUC exceeds 90%. NN was the best classifier when paired with the Chi-Square feature reduction technique, with accuracies of 87.04% and 88.32% on the training and testing sets, respectively. Meanwhile, the logistic regression classifier showed an average increase in accuracy of 8.1% during training and testing when combined with the Information Gain and ReliefF filter techniques. On the other hand, Table 8 revealed an accuracy result using the kNN-LOF outlier detection technique of 77.54%, which is lower by 9.5% than the result of the IsFLOF outlier detection technique.

The experimental results in Tables 9–13 in this study relate to resampling to reduce prediction bias from setting imbalanced sample classes using the Random resampling techniques. Fig. 4 shows the 100% resampling result data in the primary dataset of dengue infection cases and those mined using the

Table 9. High accuracy results of 3 significant classifications in the resample primary dataset of dengue infection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	ReliefF	0.985	0.859	0.858	0.857	0.859	0.989	0.867	0.865	0.864	0.867
Decision Tree	FCBF	0.866	0.799	0.797	0.796	0.799	0.865	0.799	0.795	0.793	0.799
K-Nearest Neighbors	Information Gain	0.912	0.756	0.752	0.750	0.756	0.920	0.768	0.768	0.769	0.768

Table 10. High accuracy results of 3 significant classifications in the resample dataset of dengue infection using IsF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	Chi-Square	0.986	0.896	0.895	0.895	0.896	0.914	0.911	0.911	0.910	0.911
Decision Tree	ReliefF	0.916	0.852	0.849	0.847	0.852	0.922	0.859	0.856	0.855	0.859
K-Nearest Neighbors	Information Gain	0.941	0.805	0.798	0.796	0.805	0.942	0.808	0.801	0.798	0.808

Table 11. High accuracy results of 3 significant classifications in the resample dataset of dengue infection using LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.983	0.912	0.911	0.911	0.912	0.917	0.916	0.955	0.953	0.957
Decision Tree	Chi-Square	0.915	0.859	0.858	0.856	0.859	0.919	0.860	0.857	0.856	0.860
K-Nearest Neighbors	Chi-Square	0.939	0.802	0.797	0.795	0.802	0.941	0.804	0.797	0.796	0.804

Table 12. High accuracy results of 3 significant classifications in the resample dataset of dengue infection using IsFLOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.969	0.935	0.923	0.922	0.925	0.976	0.951	0.933	0.932	0.935
Decision Tree	Chi-Square	0.963	0.904	0.902	0.900	0.904	0.971	0.905	0.902	0.901	0.905
K-Nearest Neighbors	FCBF	0.959	0.897	0.894	0.893	0.897	0.962	0.898	0.895	0.893	0.898

Table 13. High accuracy results of 3 significant classifications in the resample dataset of dengue infection using kNN-LOF

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.963	0.874	0.873	0.873	0.874	0.963	0.877	0.873	0.873	0.874
AdaBoost	ReliefF	0.964	0.868	0.867	0.866	0.868	0.964	0.868	0.867	0.866	0.868
Random Forest	Information Gain	0.962	0.867	0.865	0.865	0.867	0.962	0.867	0.865	0.865	0.867

IsF, LOF, IsFLOF, and kNN-LOF outlier detection methods.

Table 9 demonstrates that the AdaBoost classifier model and ReliefF feature selection achieved 85.9% and 86.7% accuracy in training and testing on the dengue infection cases dataset, respectively. The classifier models with the highest accuracy, listed sequentially in order, are decision tree, KNN, neural network, logistic regression, NB, random forest, and SVM. These accuracy results aligned closely with the author's findings [4].

Furthermore, the classification results for the balanced dataset of dengue infection cases that have been mined using the IsF technique are shown in Table 10. In these cases, AdaBoost, integrated with the Chi-Square feature selection technique, realized the highest accuracy of 89.6% in the training set and 91.1% in the testing set. It represented a significant accuracy increase of 3.72% compared to the resample primary dataset, which still contained outliers.

Meanwhile, the results of an experiment involving a balanced dataset mined from an outlier using the LOF technique are exhibited in Table 11. The results exhibited higher accuracy than those obtained using the IsF technique, with an improvement of approximately 1.6%. The highest accuracy was 91.2% and 91.6% in training and testing. In this case, classification on a balanced dataset cleansed from outliers using LOF outperforms rather than IsF.

In more tests, we used a balanced dataset mined using our proposed method, IsFLOF. Table 12 shows the outcomes. IsFLOF's two-layer ensemble outlier detection technique outperformed the IsF or LOF technique individually. The AdaBoost classifier, integrated with filter-based feature reduction from the FCBF technique, achieved the highest accuracy of 93.5% in training and 95.1% in testing. Thus, the proposed method shows excellence in significantly enhancing accuracy, with a 6.5% improvement compared to the primary dataset and a 7.6% improvement compared to the synthetic dataset.

Table 14. High accuracy results of 3 significant classifiers in the primary dataset of hypothyroid cases

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Decision Tree	FCBF	0.990	0.956	0.996	0.996	0.996	0.985	0.964	0.993	0.993	0.994
	Information Gain	0.974	0.952	0.992	0.991	0.992	0.993	0.963	0.993	0.993	0.994
	ReliefF	0.982	0.952	0.992	0.991	0.992	0.985	0.962	0.993	0.993	0.994

Table 15. High accuracy results of 3 significant classifications in the primary dataset of hypothyroid cases using IsF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.997	0.968	0.998	0.998	0.998	0.978	0.974	0.994	0.993	0.994
Decision Tree	Information Gain	0.991	0.966	0.996	0.996	0.996	0.991	0.972	0.991	0.991	0.992
Decision Tree	ReliefF	0.994	0.966	0.996	0.996	0.996	0.991	0.972	0.991	0.991	0.992

Table 16. High accuracy results of 3 significant classifications in the primary dataset of hypothyroid cases using LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.996	0.959	0.999	0.999	0.999	0.983	0.963	0.993	0.993	0.993
AdaBoost	ReliefF	0.993	0.957	0.997	0.997	0.997	0.983	0.962	0.992	0.992	0.992
Decision Tree	Information Gain	0.999	0.956	0.996	0.996	0.996	0.983	0.962	0.992	0.992	0.992

Table 17. High accuracy results of 3 significant classifications in the primary dataset of hypothyroid cases using IsFLOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	Chi Square	0.993	0.975	0.975	0.975	0.975	0.993	0.985	0.985	0.985	0.985
Random Forest	Information Gain	0.991	0.975	0.976	0.976	0.975	0.989	0.981	0.981	0.981	0.981
Random Forest	FCBF	0.992	0.974	0.975	0.976	0.974	0.991	0.977	0.977	0.977	0.977

Table 18. High accuracy results of 3 significant classifications in the primary dataset of hypothyroid cases using kNN-LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
Random Forest	Information Gain	0.979	0.972	0.972	0.972	0.972	0.975	0.976	0.976	0.976	0.976
Tree	Chi2	0.960	0.966	0.966	0.967	0.966	0.976	0.965	0.965	0.966	0.965
Neural Network	ReliefF	0.975	0.956	0.956	0.956	0.956	0.970	0.955	0.955	0.956	0.955

For comparison, we also tested a balanced dataset of dengue infection cases mined using kNN-LOF. Table 13 presents the results, where the highest accuracy is 87.4% and 87.7% in training and testing. These classification results are lower by 6.1% and 7.4% in training and testing compared to our proposed method, IsFLOF.

In addition, as a comparison and at the same time to verify the reliability, effectiveness, and efficiency of our proposed method, we also conducted tests on the hypothyroid dataset with identical techniques and sequences, both in the process of mining for outliers, resampling, feature selection, and classification. The original hypothyroid dataset, mined from outliers with IsF, LOF, IsFLOF, and kNN-LOF techniques, is displayed in Table 2. Meanwhile, the dataset mined and resampled is presented visually in Fig. 5.

The classification accuracy results of experiments using the original Hypothyroid dataset, and have been mined from outliers using IsF, LOF, IsFLOF, and kNN-LOF techniques, are presented in Tables 14–18. In the original dataset hypothyroid, Table 14, the Decision tree classifier excelled in all features and had an accuracy of 95.6% and 96.4% in the training and testing sets, respectively. Tables 15–18, the IsFLOF technique outperforms the kNN-LOF, ISF, and LOF strategies in classifying Hypothyroid's original dataset. Specifically, the IsFLOF technique achieves 0.3% higher accuracy than kNN-LOF, 0.7% of IsF, and 1.6% of LOF. In addition, between the IsF and LOF techniques, the IsF technique classification results are superior, with a difference of 0.9% in training and testing with AUC, F1, Precision, and Recall above 99%.

In the subsequent experiment, we used the hypothyroid dataset mined from outliers using IsF, LOF, IsFLOF, and kNN-LOF techniques and resampled using the Random resampling techniques. The results are presented in Tables 19–23. The classification results in all experiments show significant improvement, with an average accuracy above 99%. Moreover, the classification accuracy results using the IsFLOF technique outperformed all experiments using both the original dataset and the

Table 19. High accuracy results of 3 significant classifications in the resample primary dataset of hypothyroid cases

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	Information Gain	0.997	0.992	0.992	0.992	0.992	0.996	0.993	0.993	0.993	0.993
Random Forest	Chi-Square	0.996	0.989	0.989	0.989	0.989	0.997	0.985	0.985	0.985	0.985
AdaBoost	FCBF	0.998	0.988	0.988	0.988	0.988	0.997	0.988	0.988	0.988	0.988

Table 20. High accuracy results of 3 significant classifications in the resample dataset of hypothyroid cases using IsF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.998	0.996	0.996	0.997	0.998	0.997	0.997	0.998	0.997	0.996
	Information Gain	0.997	0.996	0.995	0.996	0.996	0.996	0.995	0.995	0.996	0.995
	Chi-Square	0.997	0.995	0.993	0.997	0.995	0.997	0.994	0.996	0.995	0.995

Table 21. High accuracy results of 3 significant classifications in the resample dataset of hypothyroid cases using LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	Information Gain	0.998	0.997	0.998	0.997	0.998	0.999	0.998	0.998	0.998	0.998
	ReliefF	0.996	0.997	0.998	0.996	0.998	0.998	0.996	0.996	0.996	0.997
	Chi-Square	0.997	0.996	0.996	0.996	0.996	0.998	0.997	0.998	0.997	0.996

Table 22. High accuracy results of 3 significant classifications in the resampled dataset of hypothyroid cases using IsFLOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	FCBF	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	Chi-Square	0.998	0.998	0.998	0.998	0.999	0.999	0.998	0.999	0.998	0.999
	Information Gain	0.998	0.997	0.998	0.997	0.998	0.998	0.998	0.998	0.997	0.998

Table 23. High accuracy results of 3 significant classifications in the resample dataset of hypothyroid using kNN-LOF detection

Model/Algorithms	Feature Selection	TRAINING SETS					TESTING SETS				
		AUC	CA	F1	Prec.	Rec.	AUC	CA	F1	Prec.	Rec.
AdaBoost	Information Gain	0.997	0.996	0.996	0.996	0.996	0.998	0.997	0.997	0.997	0.997
AdaBoost	ReliefF	0.998	0.996	0.996	0.996	0.996	0.998	0.997	0.997	0.997	0.997
Random Forest	FCBF	0.997	0.998	0.998	0.998	0.998	0.999	0.999	0.999	0.999	0.999

dataset mined using the kNN-LOF technique and IsF and LOF individually. The AdaBoost classifier combined with the FCBF feature selection technique on the hypothyroid resample dataset has high accuracy, AUC, F1, Precision, and Recall of 99.9% on both training and testing sets. In contrast, the classification results using a kNN-LOF detection technique on hypothyroid balanced data have lower accuracy than IsFLOF at 0.3%.

Based on the experimental results we have completed, which are presented in Tables 4-23, they demonstrate the effectiveness of our proposed new method approach. We compared the classification results for the original dengue infection and hypothyroid case datasets. This comparison also involved the application of outlier detection techniques such as IsF, LOF, IsFLOF, and kNN-LOF. In addition, we also compared the classification results on a balanced dataset using random resampling and feature selection filter-based techniques. The results indicated a significant improvement in prediction accuracy in each classification investigation. We then use these experimental results to conclude that our proposed new method approach has the consistent and most optimal accuracy in predicting dengue infection, which is 93.5% and 95.1% on the training and testing sets, respectively. The accuracy is higher than the study of Fahmi et al. [3,4] by 72.4% and 86.7% and Mello-Román et al. [32] at 72% in predicting arboviral diseases, especially dengue infection cases. There is a remarkably significant increase in accuracy in our proposed new method approach, which is 22.7%, 8.4%, and 23.1%. The accuracy achieved in this study surpasses the experimental results of Cheng et al. [11], who utilized IsF, LOF, and their combination approaches. Their combination of IsF and LOF on synthetic datasets produced an accuracy of 98%, while on real-world datasets, the accuracy was 72%. To provide a contrast, we also evaluated the accuracy of our proposed technique in the classification process using the hypothyroid case dataset. The accuracy was 99.9% on the training and

Table 24. Differences in accuracy result in the original data, IsF, LOF, IsFLOF, and kNN-LOF.

	Primary dataset	IsF	LOF	IsFLOF	kNN-LOF
Dataset of dengue infection cases					
Training	0.720	0.830	0.824	0.870	0.775
Testing	0.724	0.841	0.834	0.883	0.778
Balanced dataset of dengue infection cases					
Training	0.859	0.896	0.912	0.935	0.874
Testing	0.867	0.911	0.916	0.951	0.877
Dataset of hypothyroid cases					
Training	0.956	0.968	0.959	0.975	0.972
Testing	0.964	0.974	0.963	0.985	0.976
Balanced dataset of hypothyroid cases					
Training	0.992	0.996	0.997	0.999	0.996
Testing	0.993	0.997	0.998	0.999	0.997

testing sets. These accuracy results are also higher by 0.34% and 0.9% compared to the results obtained by Guleria et al. [33] and Chaganti et al. [34] of 99.56% and 99%.

Finally, to justify the superiority of our proposed method approach, we evaluate it and compare it with the state-of-the-art method introduced by Xu et al. [12], specifically kNN-LOF, based on the AUC metric. The highest AUC value obtained by the kNN-LOF algorithm on the dengue infection case dataset is 96.4% and 99.8 on the hypothyroid dataset. In contrast, our proposed method achieves an AUC value of 96.9% on the dengue infection case dataset and 99.9 on the hypothyroid case. Our proposed new method approach, IsFLOF, surpasses the accuracy and AUC values of kNN-LOF. This AUC comparison can be clearly seen in Tables 12, 13, 22 and 23.

Table 24 summarizes the highest classification results of our proposed new method approach, IsFLOF, compared with IsF, LOF, and kNN-LOF outlier detection techniques for predicting dengue infection and hypothyroidism cases. Table 24 explicitly confirms that mining outliers using IsFLOF as adopting the concepts of IsF, LOF, and their combination [11,13], resampling, and reduction of insignificant features can solve the classification problem on imbalanced real-world datasets measured accuracy, AUC, F1, precision, and recall an average of 95.1% in dengue infection cases and 99.9% in hypothyroid.

5. Conclusion

Improving dataset quality by reducing noise, eliminating outlier data, balancing the skewed data class instances, and selecting significant features during the data preprocessing stage proved remarkably influential in enhancing prediction accuracy on imbalanced datasets of real-world dengue infection cases and hypothyroidism. A two-layer ensemble technique called IsFLOF, which involves isolation forest (IsF) and local outlier factor (LOF), proved very efficient and effective for mining outliers and reducing complexity compared to IsF and LOF techniques individually and kNN-LOF. The proposed technique approach shows more accurate conclusions in this research area, especially in improving prediction accuracy compared to the original and synthetic datasets.

Further research is needed to improve the performance of classifiers in increasing accuracy using several approaches in outlier mining, resampling, and feature selection, such as wrapper and embedded. The focus of these studies could be prediction with the best result accuracy.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Amiq Fahmi and Mauridhi Hery Purnomo; methodology, Amiq Fahmi and Diana Purwitasari; validation, Amiq Fahmi and Surya Sumpeno; software and resources, Amiq Fahmi; investigation, Amiq Fahmi and Mauridhi Hery Purnomo; data curation, Amiq Fahmi and Diana Purwitasari; writing-original draft preparation, Amiq Fahmi; writing-review and editing, Amiq Fahmi, Diana Purwitasari, Surya Sumpeno, and Mauridhi Hery Purnomo; visualization, Amiq Fahmi; supervision Surya Sumpeno; All authors read and approved the final manuscript.

Acknowledgments

This research was supported by the Faculty of Computer Science and LPPM Dian Nuswantoro University in Semarang. But also the Multimedia Computing Laboratory research group at the Faculty of Electrical Technology and Intelligent Informatics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya.

References

[1] S. K. Jha, Z. Pan, E. Elahi, and N. Patel, “A comprehensive search for expert classification methods in disease diagnosis and prediction”, *Expert Syst.*, Vol. 36, No. 1, pp. 1–35, 2019.

[2] S. R. S. D. Neto, T. T. Oliveira, I. V. Teixeira, S. B. A. D. Oliveira, V. S. Sampaio, T. Lynn, P. T. Endo, “Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review”, *PLoS*

Negl. Trop. Dis., Vol. 16, No. 1, p. e0010061, 2022.

[3] A. Fahmi, D. Purwitasari, S. Sumpeno, and M. H. Purnomo, "Performance Evaluation of Classifiers for Predicting Infection Cases of Dengue Virus Based on Clinical Diagnosis Criteria", In: *Proc. of 2020 International Electronics Symposium (IES)*, 2020, pp. 456–462.

[4] A. Fahmi, F. A. Muqtadiroh, D. Purwitasari, S. Sumpeno, and M. H. Purnomo, "A Multi-Class Classification of Dengue Infection Cases with Feature Selection in Imbalanced Clinical Diagnosis Data", *Int. J. Intell. Eng. Syst.*, p. 17, 2022, doi: 10.22266/ijies2022.0630.15.

[5] V. J. Hodge and J. Austin, "An evaluation of classification and outlier detection algorithms", *ArXiv Prepr. ArXiv180500811*, 2018.

[6] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier Detection: Methods, Models, and Classification", *ACM Comput. Surv.*, Vol. 53, No. 3, p. 55:1-55:37, 2020.

[7] P. N. Jyothi, D. R. Lakshmi, and K. V. S. N. R. Rao, "A Supervised Approach for Detection of Outliers in Healthcare Claims Data", *J. Eng. Sci. Technol. Rev.*, Vol. 13, No. 1, pp. 204–214, 2020.

[8] B. Tarle and M. Akkalakshmi, "Integrating Multiple Techniques to Enhance Medical Data Classification", In: *Designing User Interfaces With a Data Science Approach*, IGI Global, pp. 252–274, 2022.

[9] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods", *J. Big Data*, Vol. 7, No. 1, p. 52, 2020.

[10] C. C. Aggarwal and S. Sathe, *Outlier Ensembles*. Cham: Springer International Publishing, 2017.

[11] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier factor", In: *Proc. of the Conference on Research in Adaptive and Convergent Systems*, in RACS '19. New York, NY, USA: Association for Computing Machinery, pp. 161–168, 2019.

[12] H. Xu, L. Zhang, P. Li, and F. Zhu, "Outlier detection algorithm based on k-nearest neighbors-local outlier factor", *J. Algorithms Comput. Technol.*, Vol. 16, pp. 1-12, 2022.

[13] R. Alsini, A. Almakrab, A. Ibrahim, and X. Ma, "Improving the outlier detection method in concrete mix design by combining the isolation forest and local outlier factor", *Constr. Build. Mater.*, Vol. 270, p. 121396, 2021.

[14] C. K. Aridas, S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Random Resampling in the One-Versus-All Strategy for Handling Multi-class Problems", In: Boracchi, G., Iliadis, L., Jayne, C., Likas, A. (eds), "Engineering Applications of Neural Networks". EANN 2017, *Communications in Computer and Information Science*, vol 744, pp. 111–121, 2017.

[15] F. Charte, A. J. Rivera, M. J. D. Jesus, and F. Herrera, "Addressing Imbalance in Multilabel Classification: Measures and Random Resampling Algorithms", *Neurocomputing*, Vol. 163, pp. 3–16, 2015.

[16] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection", *Int. J. Comput. Digit. Syst.*, Vol. 10, No. 1, pp. 277–286, 2021.

[17] M. Cherrington, F. Thabtah, J. Lu, and Q. Xu, "Feature Selection: Filter Methods Performance Challenges", In: *Proc. of 2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–4, 2019.

[18] H. Djellali, S. Guessoum, N. G. Zine, and S. Layachi, "Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection", In: *Proc. of 2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B)*, pp. 1–6, 2017.

[19] E. Ibrahim, M. A. Shouman, H. Torkey, and A. El-Sayed, "Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system", *Multimed. Tools Appl.*, Vol. 80, No. 13, pp. 20125–20147, 2021.

[20] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction", *Artif. Intell. Med.*, Vol. 104, p. 101815, 2020.

[21] S. Jaiswal, R. Brindha, and S. Lakhotia, "Credit Card Fraud Detection Using Isolation Forest and Local Outlier Factor", *Ann. Romanian Soc. Cell Biol.*, pp. 4391–4396, 2021.

[22] S. C. SR and H. Rajaguru, "Effective Breast Tumor Classification using K-Strongest Strength With Local Outlier Factor Algorithm", *Int. J. Aquat. Sci.*, Vol. 12, No. 03, 2021.

[23] A. B. Yusuf, R. M. Dima, and S. K. Aina, "Optimized Breast Cancer Classification using Feature Selection and Outliers Detection", *J. Niger. Soc. Phys. Sci.*, pp. 298–307, 2021.

[24] M. Bongaerts, P. Kulkarni, A. Zammit, R. Bonte, L. A. J. Kluijtmans, H. J. Blom, U. F. H. Engelke, D. M. J. Tax, G. J. G. Ruijter, and M. J. T. Reinders, "Benchmarking Outlier Detection Methods for Detecting IEM Patients in

Untargeted Metabolomics Data", *Metabolites*, Vol. 13, No. 1, Art. No. 1, 2023.

[25] A. Smiti, "A critical overview of outlier detection methods", *Comput. Sci. Rev.*, Vol. 38, p. 1-11, 2020.

[26] A. Jalalifar, H. Soliman, M. Ruschin, A. Sahgal, and A. S. Naini, "A Brain Tumor Segmentation Framework Based on Outlier Detection Using One-Class Support Vector Machine", In: *Proc. of 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1067–1070, 2020.

[27] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms", *WIREs Data Min. Knowl. Discov.*, Vol. 8, No. 6, p. e1280, 2018.

[28] S. Sugidamayatno and D. Lelono, "Outlier Detection Credit Card Transactions Using Local Outlier Factor Algorithm (LOF)", *IJCCS Indones. J. Comput. Cybern. Syst.*, Vol. 13, No. 4, Art. No. 4, 2019.

[29] R. Gao, T. Zhang, S. Sun, and Z. Liu, "Research and improvement of isolation forest in detection of local anomaly points", In: *Proc. of Journal of Physics: Conference Series*, p. 052023, 2019.

[30] M. A. Zarif and J. Hamidzadeh, "Improving performance of multi-label classification using ensemble of feature selection and outlier detection", In: *Proc. of 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 073–079, 2022.

[31] F. Thabtah, F. Kamalov, S. Hammoud, and S. R. Shahamiri, "Least Loss: A simplified filter method for feature selection", *Inf. Sci.*, Vol. 534, pp. 1–15, 2020.

[32] J. D. M. Román, J. C. M. Román, S. G. Guerrero, and M. García-Torres, "Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay", *Comput. Math. Methods Med.*, Vol. 2019, p. e7307803, 2019.

[33] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, "Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning", *Meas. Sens.*, Vol. 24, p. 100482, 2022.

[34] R. Chaganti, F. Rustam, I. D. L. T. Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques", *Cancers*, Vol. 14, No. 16, p. 3914, 2022.

[35] R. I. Kemenkes, *Pedoman pencegahan dan pengendalian demam berdarah dengue di Indonesia*, Jakarta, 2017.

[36] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams", *Big Data Cogn. Comput.*, Vol. 5, No. 1, Art. No. 1, 2021.

[37] E. Frank, "Oversampling and Undersampling," WEKA Blog. Accessed: Dec. 01, 2021. [Online]. Available: https://waikato.github.io/ weka-blog/posts/2019-01-30-sampling/

[38] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. B. Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection", *IEEE Access*, Vol. 8, pp. 132911–132921, 2020.

[39] Y. Zhang, X. Ren, and J. Zhang, "Intrusion detection method based on information gain and ReliefF feature selection", In: *Proc. of 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–5, 2019.

[40] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", *J. King Saud Univ.-Comput. Inf. Sci.*, Vol. 29, No. 4, pp. 462–472, 2017.

[41] X. Deng, M. Li, L. Wang, and Q. Wan, "RFCBF: enhance the performance and stability of Fast Correlation-Based Filter", *Int. J. Comput. Intell. Appl.*, Vol. 21, No. 02, p. 2250009, 2022.

[42] A. Salam, S. S. Prasetiyowati, and Y. Sibaroni, "Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest", *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, Vol. 4, No. 3, pp. 531–536, 2020.

[43] A. Fahmi, E. Sugiarto, A. Winarno, S. Sumpeno, and M. H. Purnomo, "Waqf Lands Assets Classification Based On Productive Value For Business Development Using Naïve Bayes", In: *Proc. of 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 622–626, 2018.

[44] R. R. Waliyansyah and U. H. A. Hasbullah, "Comparison of Tree Method, Support Vector Machine, Naïve Bayes, and Logistic Regression on Coffee Bean Image", *Emit. Int. J. Eng. Technol.*, Vol. 9, No. 1, Art. No. 1, 2021.

[45] L. Tanner, M. Schreiber, J. G. H. Low, A. Ong, T. Tolfvenstam, Y. L. Lai, L. C. Ng, Y. S. Leo, L. T. Puong, S. G. Vasudevan, C. P. Simmons, M. L. Hibberd, and E. E. Ooi, "Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness", *PLoS Negl. Trop. Dis.*, Vol. 2, No. 3, p. e196, 2008.

[46] N. Zhao, K. Charland, M. Carabali, E. O. Nsoesie, M. M. Giroux, E. Rees, M. Yuan, C. G.

Balaguera, G. J. Ramirez, and K. Zinszer, "Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia", *PLoS Negl. Trop. Dis.*, Vol. 14, No. 9, p. e0008056, 2020.

[47] Y. Cao, Q. G. Miao, J. C. Liu, and L. Gao, "Advance and Prospects of AdaBoost Algorithm", *Acta Autom. Sin.*, Vol. 39, No. 6, pp. 745–758, 2013.