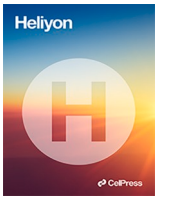
Heliyon 10 (2024) e26977

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Identification of self-care problem in children using machine learning

Maya John [a,*], Hadil Shaiba [b]

[a] Artificial Intelligence and Data Analytics (AIDA) Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia

[b] Department of Computer Science, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

ARTICLE INFO

Keywords:
Self-care problem
machine learning
Imbalanced data
Classification

ABSTRACT

Identification of self-care problems in children is a challenging task for medical professionals owing to its complexity and time consumption. Furthermore, the shortage of occupational therapists worldwide makes the task more challenging. Machine learning methods have come to the aid of reducing the complexity associated with problems in diverse fields. This paper employs machine learning based models to identify whether a child suffers from self-care problems using SCADI dataset. The dataset exhibited high dimensionality and imbalance. Initially, the dataset was converted into lower dimensionality. Imbalanced dataset is likely to affect the performance of machine learning models. To address this issue, SMOTE oversampling method was used to reduce the wide variations in the class distribution. The classification methods used were Naïve bayes, J48 and random forest. Random forest classifier which was operated on SMOTE balanced data obtained the best classification performance with balanced accuracy of 99%. The classification model outperformed the existing expert systems.

1. Introduction

As per the World Health Organization (WHO), disability is considered as a conglomerate of impairments, limited activities and restricted participation [1]. International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY) is a framework developed by WHO for classifying disability among children up to 18 years of age [2]. The aforementioned framework may be used in diverse fields such as medicine, formulation of policies, monitoring of children etc. Disability can be divided broadly into physical, motor and mental. The most common types of disability in children are physical and motor disability. These disabilities prevent children from caring about their daily chores without support from others [3]. It has been reported that identification of disability is a tough task, and it requires the support from medical professionals in particular occupational therapists [4]. COVID-19 pandemic has increased the demand for occupational therapists as the viral infection caused a lot of people to experience muscular, skeletal, neurological issues even after recovery. There is wide gap between the demand and availability of occupational therapists in several countries. Hence, there arises the need for developing computer aided methods to ease the work of identifying disabilities [5].

Identification of self-care problem in children is highly essential as properly dealing with this type of problem is instrumental in ensuring good quality of life. Examples for self-care problems include not being able to care for one's body parts, toileting issues, eating

* Corresponding author.
E-mail addresses: mjohn@psu.edu.sa (M. John), hashaiba@pnu.edu.sa (H. Shaiba).

https://doi.org/10.1016/j.heliyon.2024.e26977
Received 16 September 2023; Received in revised form 14 February 2024; Accepted 22 February 2024
Available online 28 February 2024

issues, self-safety issues, getting dressed problem etc. This paper proposes a random forest based classifier to identify whether a child suffers from self-care problem. The present work uses SCADI (Self-Care Activities Dataset based on ICF-CY) dataset. Prior to classifying children on the basis of self-care problem, the dataset was pre-processed as it had large number of features and highly uneven class distribution. The remainder of the paper is structured as follows: section two presents the survey of similar works, section three deals with the data and methods used, the results obtained and discussion is described in section four and section five concludes the paper.

2. Literature survey

Imbalance in a dataset may reduce the effectiveness of the classifier [6]. A number of publications have stated that oversampling data using SMOTE enhanced the classifier performance. This oversampling method has proved its effectiveness in variety of applications. A research work on predicting hypertension and type 2 diabetes, employed SMOTE to improve its efficiency [7]. The paper employed Density-based Spatial Clustering of Applications (DBSCAN) to detect outliers in data, SMOTE for balancing the data and random forest for classifying the data. SMOTE based expert system attained high accuracy in predicting heart failure [8]. Decision Table/Naive Bayes hybrid classifier which operated on SMOTE modified data was most effective in predicting whether a person would have heart failure. SMOTE has also been employed to balance data related to credit risk prediction [9]. SMOTE technique helped in developing a system with enhanced colon cancer prediction accuracy [10]. Random forest and XGBoost based classifier attained better results than logistic regression based classifier.

In the past, several machine learning based expert systems have been developed to help medical professionals in classifying self-care problem in children. Syafrudin et al. created a novel model using heuristic method (genetic algorithm) and extreme gradient boosting (XGBoost) to classify children's self-care issues. The most appropriate features were chosen using genetic algorithm [11]. Nearest neighbour-based classifier operating on feature set reduced using principal component analysis was found to be effective in identifying the category of self-care problems [12]. Probabilistic neural network [13], deep forward neural network [14], etc. have been used for multi class categorization of self-care issues.

The works mentioned in the previous paragraph deals with identifying the type of self-care problem a child is facing. Very few works associated with expert systems that are based on self-care problems deal with identifying whether a child has self-care problem

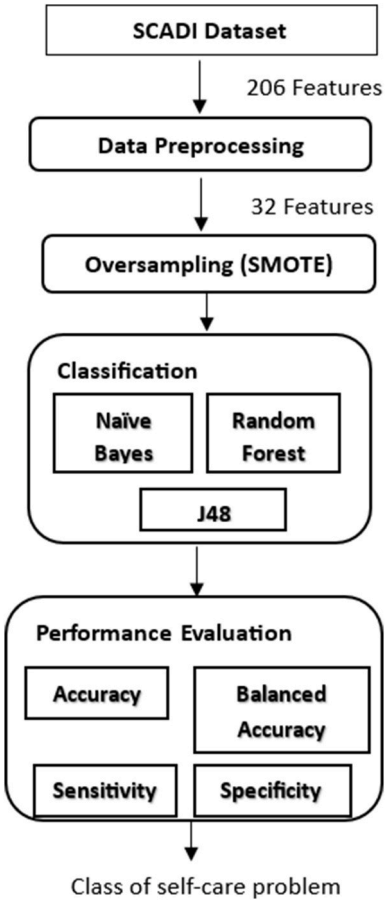


Fig. 1. Overview of research work.

or not. Hence there is a scope for developing such systems with better effectiveness. Correctly identifying the children with self-care issues helps in providing them with proper care/training to enhance their quality of living. The genetic algorithm-XGBoost hybrid method classified data about children with self-care issues accurately in 98.57% of the cases [11]. Putatunda employed techniques such as autoencoder and deep neural network to perform binary classification on self-care dataset [15]. The system developed attained an accuracy of around 91%.

3. Data and methods

The overview of the research work is depicted in Fig. 1.

3.1. Dataset

The research used freely accessible SCADI (Self-Care Activities Dataset) dataset. The dataset developed by Zarchi et al., consists of data pertaining to the self-care problems faced by children disabled in terms of physical and motor activities [16]. The dataset consists of 206 features. The features considered include gender, age, 203 attributes related to self-care activities and the category of self-care problem (target attribute). The activities considered may be broadly classified as related to washing oneself, taking care of one's body parts, sanitation related, dressing, consumption of solid and liquid foods, attending to one's health and safety, etc. A total of 29 self-care activities were considered and seven features were used to describe each activity. The seven features corresponding to each activity were related to the level of difficulty a child faced with regard to the activity. These binary features were no disability, slight disability, moderate disability, significant disability, complete disability, not applicable and not specified. The original dataset consists of details of 70 children and seven categories of self-care issues. In the present work, we have considered the problem as a two class classification task with one class corresponding to no problem instances and the other class dealing with instances having self-care issues.

3.2. Data pre-processing

It was observed that out of the seven binary features used to describe each self-care activity, only any one feature has value one while the rest six features have the value zero. Hence the seven features were replaced by one attribute which indicated the degree of level of difficulty a child faced regarding the activity. Therefore 203 features corresponding to self-care activities were reduced to 29 features. Hence the modified dataset consisted of just 32 attributes (that is: gender, age, 29 features related to self-care activities and class of self-care problem).

3.2.1. SMOTE

Out of the 70 instances in the dataset, 14 instances correspond to children with no self-care problems, that shows that the dataset is imbalanced as three fourths of the instances suffer from self-care problems. Hence, there is a chance that the performance of the classifiers may be biased towards the majority class. Various approaches, such as under-sampling, over-sampling, hybrid sampling etc. have been employed in the past to handle imbalanced data effectively. In this paper, Synthetic Minority Oversampling Technique (SMOTE) is employed to deal with the disparity in the class distribution. The percentage of oversampling was set at 250% and the oversampling was carried out based on five nearest neighbors.

SMOTE is an oversampling method wherein new samples are created based on k-closest instances, while typical over-sampling techniques create new samples that are a replication of the existing samples [17]. The value of k needs to be specified along with the percentage of over-sampling that is required. For example, if $k = 2$ and the amount of over-sampling is 150%, two nearest neighbors are chosen, and two samples designated *sample*1 and *sample*2 are then randomly chosen from the neighbors to create a new sample. The idea is to generate a new sample that is close to the randomly selected minority samples by using equations (1) and (2):

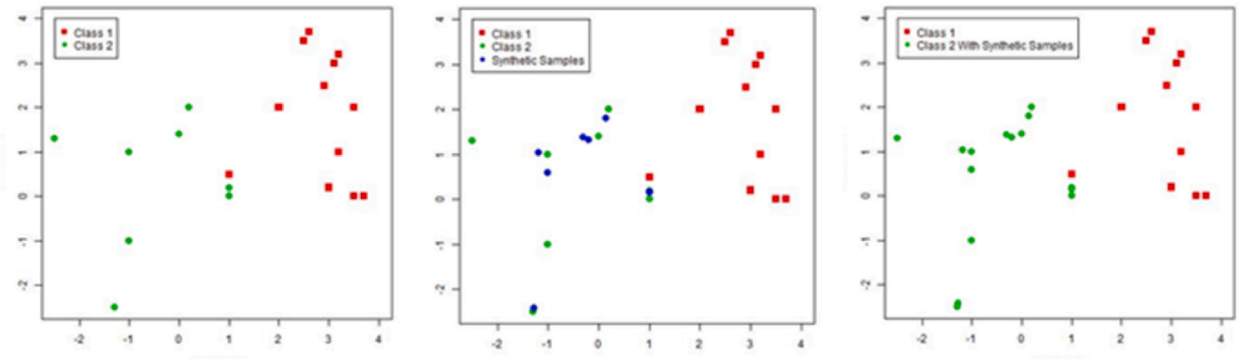


Fig. 2. An example of SMOTE with k = 2.

$$\text{difference} = \text{sample2-sample1} \tag{1}$$

$$\text{synthetic} = \text{sample1} + (\text{rand} * \text{difference}) \tag{2}$$

where *difference* is the difference between the two randomly selected samples, *rand* is a randomly selected number in the range 0–1 and *synthetic* is the newly created sample.

An illustration of SMOTE with $k = 2$ (two nearest neighbors) is shown in Fig. 2.

3.3. Cross-validation

Our study uses stratified k-fold cross-validation (CV). The technique is effective in dealing with underfitting/overfitting which may occur when dealing with an imbalanced dataset [18]. In k-fold CV, the data is split into k groups also known as folds. One-fold is used for testing the model and the rest for training. This process is repeated k times with different folds being considered as test data. The performance of the model is computed as the average of performance of model in each step. Unlike k-fold CV, stratified k-fold CV ensures that the classes are evenly distributed in different folds.

3.4. Classification

The different classification techniques used for identifying whether a child has self-care problem is described in the following subsections.

3.4.1. J48

Decision trees are one of the most effective supervised learning algorithms which have demonstrated excellent efficacy in an array of applications due to its numerous benefits. Decision trees have a number of benefits, including the ability to handle data with missing values, the ability to be applied to various data types (continuous and categorical), the simplicity of its structure which makes it easy to understand and explain and the capacity to operate well with small datasets. A decision tree comprises of root node, internal nodes and leaf nodes. The decision classes are located in the leaf nodes, while the test conditions are contained in the internal nodes. Univariate decision trees use a single attribute for decision making in their internal nodes. Two examples of such a tree are ID3 and J48. These trees make use of key concepts like information gain, entropy and tree pruning. In J48, the information gain of features is primarily used to choose the attributes for splitting the node, the attributes with maximum values of information gain are selected. Information gain is computed using the measure of differences in entropy, which is a metric for how random a feature is, so higher uncertainty levels are linked to low entropy levels.

3.4.2. Random forest

Random forest is a higher version of decision trees which classifies data based on predictions of multiple decision trees [19]. The class predicted by a random forest is based on the prediction of maximum number of the decision trees. The core concepts behind the working of random forest are bootstrap aggregation (bagging) and random feature selection. Bagging employs an ensemble of classifiers to enhance the accuracy of weak learner classifier. Fig. 3 is an illustration of random forests.

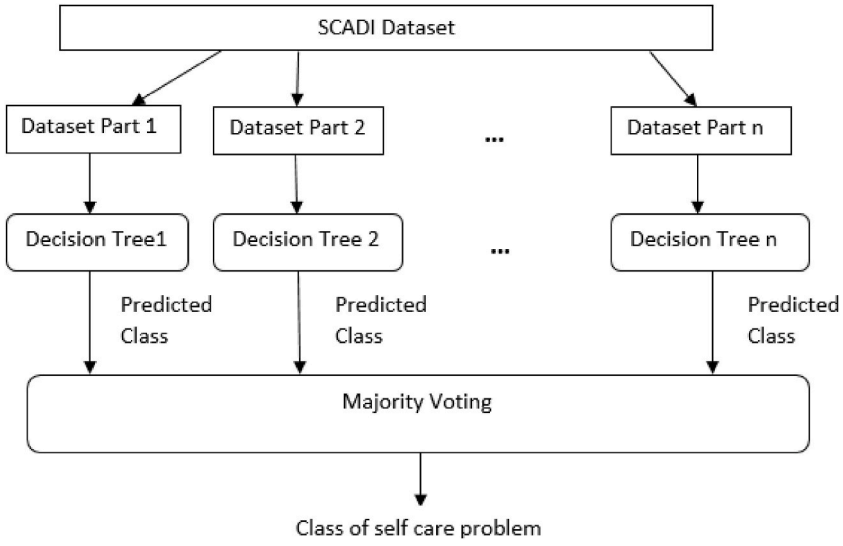


Fig. 3. An illustration of random forest.

4. Naïve Bayes

The basis of naïve bayes-based classifier is the Bayesian theorem. The classifier ignores the dependencies between the input variables [20]. Each input feature is equally important in determining the output variable. The mathematical formulation of Bayesian theorem is as shown in Equation (3).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

Where A and B are events. Here probability of event A occurring is computed provided that event B has already occurred. Suppose the output variable is considered as y and $x_1, x_{2\ldots}\ x_n$ are the input variables. Then the conditional probability of y depends on x values is computed as shown in equation (4).

$$P(y|x_a, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \tag{4}$$

5. Results and discussion

The experiments were performed using Weka on the modified dataset which consists of 32 features after data pre-processing. Stratified five-fold CV strategy was adopted to evaluate the classifiers.

The effectiveness of different classifiers were compared using metrics such as sensitivity, specificity, accuracy and balanced accuracy. In this work, the children with self-care problems constitute the positive class and children with no self-care problems come under the negative class.

Accuracy refers to the portion of samples classified correctly and is computed according to equation (5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Sensitivity refers to the portion of positive class samples classified correctly. It is computed as per equation (6).

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

Specificity refers to the ratio of number of negative class samples classified accurately to the total number of negative class samples. It is calculated as shown in equation (7).

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

Imbalance in data causes the model to be biased towards the majority class. Hence, high accuracy value is achieved even if very few instances of the minority class are classified accurately. Balanced accuracy is an evaluation metric which gives equal importance to classification accuracy of classes with majority number of samples as well as minority number of samples and hence is considered as an effective metric of measuring the performance of classifiers handling imbalanced data. Balanced accuracy is calculated as per equation (8).

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{8}$$

The default parameters were used for classification using Naïve bayes, J48 and random forest (for feature excluding number of trees). On experimenting with AutoWeka, it was observed that random forest gave the best performance when the number of trees was set to 10. Hence the number of trees in case of random forest was set to 10. The performance of the classifiers on the original and modified dataset are tabulated in Tables 1 and 2 respectively.

It is evident from Table 1 that in the case of the original dataset, the highest specificity value was attained by the naïve bayes classifier and random forest was most effective in classifying the positive class instances (indicated by high sensitivity value). The dataset was modified as described in section 3.2. It can be observed from Table 2 that in the case of modified dataset, naïve bayes and J48 classifier were most effective in classifying the positive instances while random forest classifier obtained the best specificity.

The comparison between the different classifiers in terms of sensitivity is shown in Fig. 4. It is highly essential for classifiers to attain high sensitivity as it refers to accuracy in correctly predicting the instances related to children with self-care problems. It can be

Table 1
Performance of different classifiers on the original feature set.

Method	Sensitivity	Specificity
Naïve Bayes	0.91	0.88
J48	0.91	0.69
Random Forest	0.94	0.69

Table 2
Performance of different classifiers on the modified feature set.

Method	Sensitivity	Specificity
Naïve Bayes	0.96	0.75
J48	0.96	0.81
Random Forest	0.93	0.88

inferred from Fig. 4 that relatively high sensitivity is obtained in case of majority of the classifiers when the modified feature set is considered.

The comparison between accuracy and balanced accuracy of the models on the modified dataset is shown in Fig. 5. It can be inferred from the figure that there is no wide variation between the two values. Hence, it is clear that the classifiers are effective in predicting both the positive and negative class instances.

The results of the classification performance after performing oversampling on the modified dataset is tabulated in Table 3. The random forest classifier is most effective in classifying both the positive and negative classes.

The comparison of balanced accuracy of models based on data with and without SMOTE application is depicted in Fig. 6. It is evident that random forest classifier which worked on oversampled data attained very high balanced accuracy. The hybrid method using SMOTE and random forest was also effective in the prediction of high levels of blood pressure and second order diabetics [7].

Table 4 shows the comparison between the proposed system and other state-of-the-art classification techniques in terms of feature selection, classifier, number of features and accuracy. Existing systems used methods such as genetic algorithm-XGBoost [11] and autoencoder-deep neural networks [15] for identifying self-care problems in children. The proposed system showed better accuracy in classifying data when compared to existing methods. It simply used data pre-processing to reduce the number of features while the existing systems used more complicated techniques like genetic algorithm and auto encoder for feature reduction. Also, our proposed model uses relatively less number of features compared to the existing systems while performing better in terms of accuracy.

6. Conclusion

Disabilities in children leads to self-care problems which in turn adversely affects the quality of the child and his/her parents' life. Identifying this type of problem is a complex and time-consuming task. This research proposed an expert system based on random forest to identify whether a child suffers from self-care problem. Pre-processing was carried out to reduce the number of features in the dataset. As the dataset was imbalanced, SMOTE technique was used to balance the number of instances in both classes. Results of the experimentation showed that the best classifier performance was attained when the data was pre-processed and balanced. Random forest classifier attained the best accuracy value of 99.09%. These types of expert system will ease the work of medical practitioners in diagnoses and decision making. Furthermore, more analysis needs to be conducted to identify the feature importance of classifiers.

Funding statement

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R135), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data availability

The dataset used in the work was obtained from UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/446/scadi).

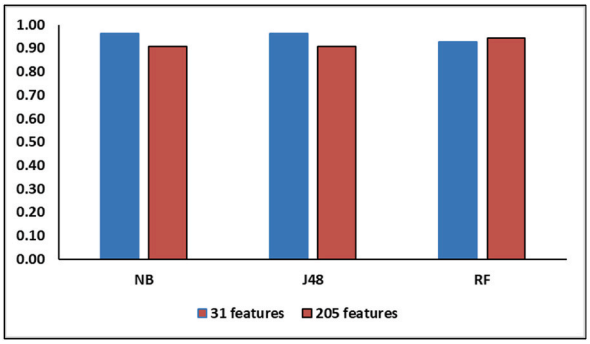


Fig. 4. Comparison of sensitivity measure based on different number of features.

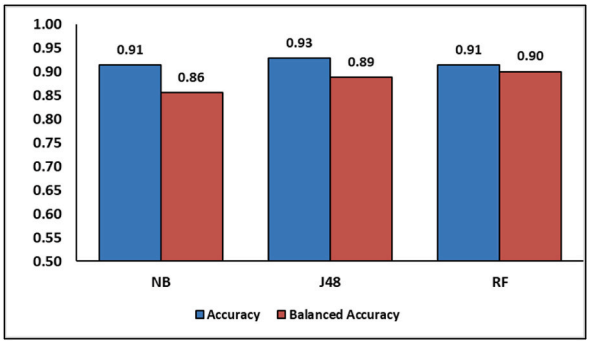


Fig. 5. Comparison between accuracy and balanced accuracy of classifying the modified feature set without SMOTE.

Table 3
Performance of different classifiers on the modified feature set after applying oversampling using SMOTE.

Method	Sensitivity	Specificity
SMOTE + Naïve Bayes	0.94	0.91
SMOTE + J48	0.94	0.96
SMOTE + Random Forest	1.00	0.98

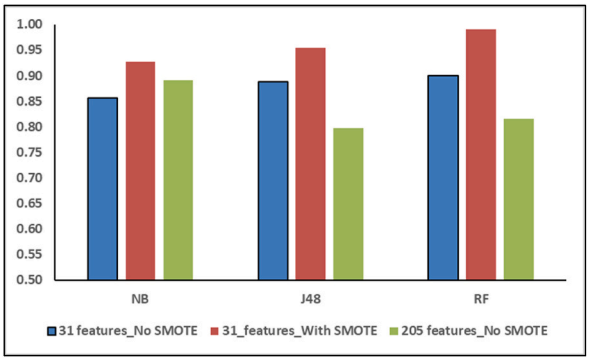


Fig. 6. Comparison between balanced accuracy of the classifiers with and without balancing the data and feature reduction.

Table 4
Proposed system versus existing works.

Method	No. of features used for classification	Accuracy
GA + XGBoost [11]	111	98.57%
Autoencoders + deep neural networks [15]	205	91.43%
Proposed System (Pre-processing + SMOTE + Random Forest)	31	99.09%

CRediT authorship contribution statement

Maya John: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hadil Shaiba:** Writing – review & editing, Validation, Project administration, Investigation, Funding acquisition, Formal analysis, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R135), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank Prince Sultan University, Riyadh, Saudi Arabia for the support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e26977.

References

[1] R. Lucas-Carrasco, E. Eser, Y. Hao, K.M. McPherson, A. Green, L. Kullmann, T.W.D. Group, The Quality of Care and Support (QOCS) for people with disability scale: development and psychometric properties, Res. Dev. Disabil. 32 (3) (2011) 1212–1225.
[2] Who, International Classification of Functioning, Disability, and Health: Children and Youth Version, ICF-CY World Health Organization, 2007.
[3] A. Dardzińska-Głębocka, M. Zdrodowska, Analysis children with disabilities self-care problems based on selected data mining techniques, Procedia Computer Science 192 (2021) 2854–2862.
[4] Y.L. Yeh, T.H. Hou, W.Y. Chang, An intelligent model for the classification of children's occupational therapy problems, Expert Syst. Appl. 39 (5) (2012) 5233–5242.
[5] T. Le, S.W. Baik, A robust framework for self-care problem identification for children with disability, Symmetry 11 (1) (2019) 89.
[6] F. Thabtah, S. Hammoud, F. Kamalov, A. Gonsalves, Data imbalance in classification: experimental evaluation, Inf. Sci. 513 (2020) 429–441.
[7] M.F. Ijaz, G. Alfian, M. Syafrudin, J. Rhee, Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest, Appl. Sci. 8 (8) (2018) 1325.
[8] S. Priyadarshinee, M. Panda, Improving prediction of chronic heart failure using SMOTE and machine learning, in: 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), IEEE, 2022, September, pp. 1–6.
[9] L. Wang, Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization, Appl. Soft Comput. 114 (2022) 108153.
[10] G. Leonard, C. South, C. Balentine, M. Porembka, J. Mansour, S. Wang, M. Augustine, Machine learning improves prediction over logistic regression on resected colon cancer patients, J. Surg. Res. 275 (2022) 181–193.
[11] M. Syafrudin, G. Alfian, N.L. Fitriyani, M. Anshari, T. Hadibarata, A. Fatwanto, J. Rhee, A self-care prediction model for children with disability based on genetic algorithm and extreme gradient boosting, Mathematics 8 (9) (2020) 1590.
[12] B. Islam, N.M. Ashafuddula, F. Mahmud, A machine learning approach to detect self-care problems of children with physical and motor disability, in: 2018 21st International Conference of Computer and Information Technology (ICCIT), IEEE, 2018, December, pp. 1–4.
[13] S.F. Bushehri, M.S. Zarchi, An expert model for self-care problems classification using probabilistic neural network and feature selection approach, Appl. Soft Comput. 82 (2019) 105545.
[14] K. Akyol, Comparing of deep neural networks and extreme learning machines based on growing and pruning approach, Expert Syst. Appl. 140 (2020) 112875.
[15] S. Putatunda, Care2Vec: a hybrid autoencoder-based approach for the classification of self-care problems in physically disabled children, Neural Comput. Appl. 32 (2020), 17669-1.
[16] M.S. Zarchi, S.F. Bushehri, M. Dehghanizadeh, SCADI: a standard dataset for self-care problems classification of children with physical and motor disability, Int. J. Med. Inf. 114 (2018) 81–87.
[17] M. John, J.S. Jayasudha, Enhancing performance of deep learning based text summarizer, Int. J. Appl. Eng. Res. 12 (24) (2017) 15986–15993.
[18] E.G. Adagbasa, S.A. Adelabu, T.W. Okello, Application of deep learning with stratified K-fold for vegetation species discrimation in a protected mountainous region using Sentinel-2 image, Geocarto Int. 37 (1) (2022) 142–162.
[19] M. John, H. Shaiba, Ensemble based foetal state diagnosis, in: 2020 6th Conference on Data Science and Machine Learning Applications (CDMA), IEEE, 2020, March, pp. 129–133.
[20] T.M. Mitchell, "Machine Learning", McGraw-Hill, 1997.