Grenze International Journal of Engineering and Technology, June Issue

Expert System based on Feature Selection for Classification of Self-Care Problems in Children and Youth People

Dr. V. Vijayalaskhmi[1], Dr. Saira Banu Atham [2] and Panimalar Kathiroli[3]
[1]Associate Professor/CMR Institute of Technology/ISE, Bengaluru, INDIA
Email: vijayalakshmi.v@cmrit.ac.in
[2]Program Head/Presidency University, Bengaluru, INDIA
Email: saira.atham@gmail.com
[3]Assistant Professor/Sathyabama University/CSE, Chennai, INDIA
Email: panimalar.cse@sathyabama.ac.in

***Abstract*—One of the important problems for occupational therapists is self-care problems classification of children. The International Classification of Functioning, Disability, and Health for Children and Youth (ICF-CY) is a standard language to store the characteristics of child development in terms of health, education and social activities. Two expert models were proposed using Support Vector Machine classifier and also rule based classifier to predict the self-care problems of children and youth with physical and moor disability in earlier stage. The well-known and only available SCADI (Self-Care Activities Dataset based on ICF-CY) dataset was used. This dataset consists of numerous features. Generally, too many features take more space, time and not produce good result. Dimensionality reduction was used to select relevant features which are really needed for the models. The Boruta algorithm was used for feature selection. The Support Vector Machine model was developed using whole dataset and also 48 selected features in R Programming. The model outperformed with 97.14% on selected attributes rather than all attributes. This expert system will be useful for therapist to identify the problem of exceptional children.**

***Index Terms*— SCADI dataset, Self-Care Problems, ICF-CY, Boruta algorithm, SVM classifier.**

I. INTRODUCTION

In medical sector maintaining the record of exceptional children therapy is an important work. To monitor, diagnosis and classification of those children is major challenging area. Self-care for children with physical and motor disabilities is main regular living activities which need more consideration for exceptional children. Self-care problems of children need to be classified for defining treatment by occupational therapists [8]. SCADI: A standard dataset for self-care problems classification of children with physical and motor disability is based on ICF-CY released by WHO [1, 16]. The activities like looking after one's safety, looking after one's health, caring for body parts, eating and drinking are known as self-care activities. The treatment should be chosen for a person depends on the type of problem. So, detection of self-care problem is a significant process in the health care system for exceptional children and youth. This process can be done by well experienced and professional

Grenze ID: 01.GIJET.8.2.322

therapist [2]. It is economically costly and time-consuming work. The solution was provided by developing an expert system to predict the self-care problems based on SVM.
The solution was given by researchers using machine learning algorithms such as ANN [1], KNN [2], RF [3,7], MLP [4], FNN [6], GA+PNN [8], FSX [9], ELM [10], DNN [11], Care2vec [12], GA+XGBoost [13], DT [14], NB [5,15] are utilized. Our proposed system used SCADI dataset, which consists of many features. Dimensionality reduction is helpful to reduce the number of features which are not really useful. Boruta feature selection algorithm utilized to select relevant features. 48 features confirmed attributes selected and created Support Vector Machine model. At last, the accuracies of compared with both model with all features and model with selected features. The model with selected features produced high accuracy. This paper is organized as follows; in second section the related works are given. The section three describes the proposed approach which includes data collection, model implementation and result discussion. Finally, the last section concludes the work.

II. Related Works

The standard dataset for self-care problems classification of children with physical and motor disability was provided by the authors M.S. Zarchia, S.M.M. Fatemi_Bushehrib, and M. Dehghanizadeh called SCADI (Self-Care Activities Datasets for ICF-CY) in 2018 [1]. ICF-CY means (International Classification of Functioning, Disability, and Health in Children and Youth). They developed two expert systems using Artificial Neural Networks and Rule based system using C4.5. The ANN based system provided the high accuracy 83.1% than decision tree algorithm [1].Bayezid Islam et. al., proposed an expert system to predict in advance the self-care problems of children with physical and motor disability using SCADI dataset. They used PCA to reduce the number of features. Then they applied different machine learning algorithms but KNN produced high accuracy [2]. Keleş, Mümine Kaya, and Ümit Kılıç, utilized ABC (Artificial Bee Colony) feature selection algorithm to reduce the number of features. They selected 7 from 206 features of SCADI dataset. Applied KNN, Random Forest, SMO, Decision Table and Naïve Bayes. They concluded the result increased when the features are reduced than with whole dataset [3].
Sudarshan S. Chawathe [4], presented a method for classification of self-care problems using Bayes Net, Multilayer Perceptron and JRip decision tree algorithms. Lei Liu et. al., introduced a novel feature selection technique which is based on feature curve of subclass problem called IGRCFS (Information Gain Regression Curve Feature Selection) to find the feature subset optimally. They compared IGRCFS with five different feature selection techniques with eight different datasets including SCADI. The result demonstrated that IGRCFS is best comparing to others [5]. Paulo V. C. Souza et. al., proposed intelligent hybrid systems based on fuzzy rules to develop a model to predict and analyze the disease problems in children and youngsters and obtained high accuracy [6]. Avishek Choudhury [7], proposed a predictive model to classify self-care problems on SCADI dataset based on ICF-CY. They implemented many algorithms like Random Forest, Support vector machine, Naïve Bayes, Hoeffding tree, and Lazy locally weighted learning using two-tailed T-test at 95% confidence interval. The data dimensionality reduction done by using Boruta algorithm. Random Forest produced better classification accuracy with 84.75%, RMSE with 0.18 and ROC with 0.99. S.M.M. Fatemi Bushehri a, Mohsen Sardari Zarchi [8], proposed an expert model using Genetic Algorithm (GA) and Probabilistic Neural Network (PNN) for categorizing self-care problems of kids with physical and motor disability. They used PNN as classifier and GA as feature selection method on SCADI dataset. They also extracted classification rules using CART algorithm. The results shown that the better accuracy 94.28% because of only used 16.5% features.
Tuong Le and Sung Wook Baik [9], proposed framework to improve the performance of prediction in SCADI dataset using extreme gradient boosting (FSX) and sampling technique. The FSX used the Synthetic Minority Over-sampling Technique (SMOTE) oversampling technique with different algorithms like Random Forest, Support Vector Machine, Artificial Neural Network with the same dataset. FSX produced 85.4% in accuracy for self-care problem classification in medical diagnosis.Omer Faruk Ertugrul and Necmettin Sezgin [10], investigated the classification of self-care problems using eXtreme Learning Machine (ELM) using SCADI dataset. They applied 10-fold cross validation technique and got 66% on whole features. The high accuracy 86.25% achieved when maximum relevant 81 features were selected. Kemal Akyol [11], applied Deep Neural Networks and Extreme Learning Machines architectures on two datasets such as Parkinson and SCADI dataset. The evaluation was based on growing and pruning approach. The Deep Neural Network yield better result for both datasets comparing to Extreme Learning Machines.Sayan Putatunda [12], used new deep learning approach called Care2Vec for the classification self-care problem. Care2Vec is the combination of autoencoder and deep neural network. Then they compared the result with traditional methods like decision tree and deep neural

network in both binary and multi-class classification. The mean accuracy and mean area under the curve (AUC) were used to measure the performance.

Muhammad Syafrudin et. al. [13], proposed the prediction model for self-care problems usingcombination of Genetic Algorithms and eXtreme Gradient Boosting called GA-XGBoost. GA was used as feature selection technique. They used accuracy, precision, recall and f1-score for measurements. The accuracy of GA-XGBoost was higher comparing to six other machine learning algorithms. Muhammad Syafrudin et. al, developed classification model for self-care problem for children using Principal Component Analysis (PCA) and Decision Tree (DT) [14]. The dimensionality reduction done with PCA and built model using DT with 10-fold cross-validation. The decision tree model achieved 94.29% accuracy by improving 1.7% because of limited number of features with PCA.

Agnieszka Dardzińska-Głębocka, Małgorzata Zdrodowska [15], built classification models using Naïve bayes, J48, PART, CART and Multilayer Perceptron in WEKA tool with SCADI dataset. They applied all algorithms in three different models such with all features, 39 features by chi-square test and 19 features by correlation-based feature subset selection method. Naïve Bayes achieved best accuracy in all three models. Table I depicts the survey of existing systems with methods used, accuracy with feature selection, without feature selection and feature selection technique used to create those models.

TABLE I. LITERATURE SURVEY

Author, Year	Method	Accuracy Without Feature selection (%)	Accuracy With Feature Selection (%)	Feature Selection Technique
Zarchi et al., 2018	ANN	83.1	-	-
Bayezid et al., 2018	Gradient Boosting	82.86	82.86	PCA
	Random Forest	84.29	82.86	
	ANN	82.86	82.86	
	Support Vector Machine	84.29	84.29	
	KNN	81.43	84.29	
	ELM	82.86	84.29	
Putatunda, 2018	Care2Vec	84.29	-	-
	DNN	81.43	-	-
	Decision Tree	76.99	-	
Choudhury, 2018	Random Forest	83.28	84.75	Boruta (55 Features)
	Support Vector Machine	78.86	82.43	
	Naïve Bayes	82.57	83.00	
	Lazy LWL	78.71	78.57	
	Hoeffding tree	82.57	80.71	
Kele and Keles, 2018	K Nearest Neighbor	81.43	88.57	Artificial Bee Colony
	Random forest	84.29	85.71	
	SMO	80.00	82.86	
	Decision table	77.14	87.14	
	Naïve Bayes	82.86	87.14	
Paulo et al., 2019	FNN	85.11	-	-
	C4.5	82.97	-	
	MLP	80.68	-	
	SVM	79.84	-	
Omer and Neemetin, 2019	ELM	66.00	86.25	81 features selected
Tuong and Sung, 2019	ANN	-	81.5	SMOTE over sampling
	FSX	-	85.4	
	SVM	-	79.3	
	RF	-	83.7	
Muhammad et al., 2020	DT	86.50	94.29	PCA
In this Paper	SVM	81.82	97.14	Boruta (48 Features)

III. PROPOSED APPROACH

The Figure 1 depicts the necessary steps are followed to form this model are data collection, data preprocessing, feature selection using Boruta algorithm, splitting data into training and testing data, model creation using training dataset, model evaluation using testing dataset and finally results obtained.

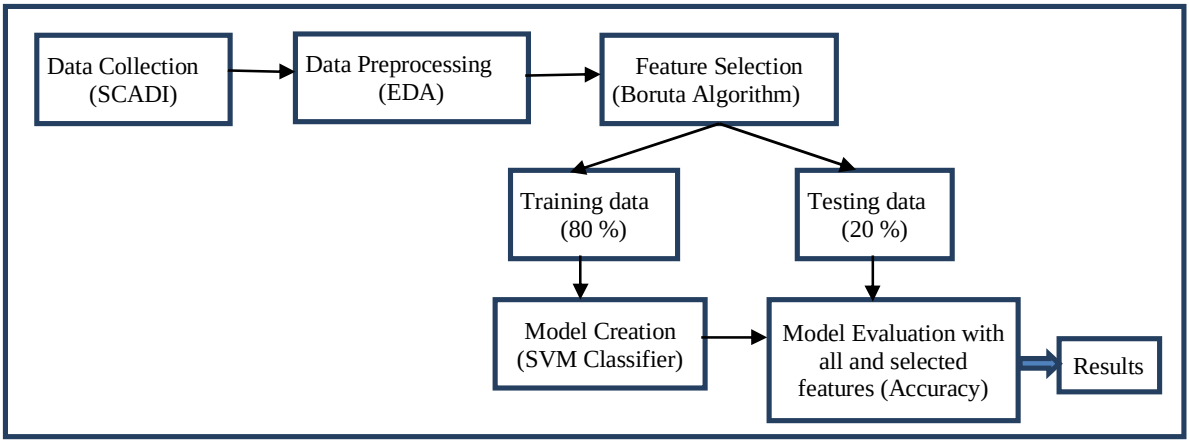

Figure 1. Architecture Diagram

A. Data Collection

The SCADI dataset describes the many functional problems of children and youth in various fields like physically, mentally, motion and biologically. To describe the health components the ICF-CY contain four levels of coding such as level 1, level 2, level 3 and level 4. The combination of letter and number called component code represents level 1. Level 2 represents numbers from 10 to 99; level 3 and 4 represents numbers ranges from 0 to 9.

Example: d53010
Level 1 = d5
Level 2= 30
Level 3 = 1
Level 4 = 0

There are 29 activities which are further divided into 8 self-care categories shown as follows in table II.

TABLE II. DESCRIPTION OF SCADI DATASET

S. No.	Category No.	Category Name	Activities	Code
1	I	Washing oneself	Washing body parts	d 5100
2	I		Washing whole body	d 5101
3	I		Drying oneself	d 5102
4	II	Caring for body parts	Caring for skin	d 5200
5	II		Caring for teeth	d 5201
6	II		Caring for hair	d 5202
7	II		Caring for fingernails	d 5203
8	II		Caring for toenails	d 5204
9	II		Caring for nose	d 5205
10	III	Toileting	Indicating need for urination	d 53000
11	III		Carrying out urination appropriately	d 53001
12	III		Indicating need for defecation	d 53010
13	III		Carrying out defecation appropriately	d 53011
14	III		Menstrual care	d 5302
15	IV	Dressing	Putting on clothes	d 5400
16	IV		Taking off clothes	d 5401
17	IV		Putting on footwear	d 5402
18	IV		Taking off footwear	d 5403
19	IV		Choosing appropriate clothing	d 5404
20	V	Eating	Indicating need for eating	d 5500
21	V		Carrying out eating appropriately	d 5501
22	VI	Drinking	Indicating need for drinking	d 5600
23	VI		Carrying out drinking appropriately	d 5601
24	VII	Look after one's health	Ensuring one's physical comfort	d 5700
25	VII		Managing diet and fitness	d 5701
26	VII		Managing medications and following health advice	d 57020
27	VII		Seeking advice or assistance from caregivers or professionals	d 57021
28	VII		Avoiding risks of abuse of drugs or alcohol	d 57022
29	VIII	Looking after one's safety	Looking after one's safety	d 571

The dataset contains 208 attributes and which was collected from 70 children with age from 6 to 18 years old (41 male, 29 female). The attributes are gender, age, 205 self-care activities and classes with 7 values as target variable. These possible values for all the features are shown in table III.

TABLE III. POSSIBLE VALUES FOR FEATURES

Gender	Age	205 self-care activities	Classes (Target Feature)
Binary Values 1 – male, 0 – female (41 male, 29 female)	Numeric Values (from 6 to 18)	Binary Values (0 – No and 1 – Yes)	Numeric Multi class values (from 1 to 7)

The number of children for each class is described in the table IV. Maximum and minimum number of children is belonging to class 6 and class 3 respectively.

TABLE IV. NUMBER OF PERSONS FOR EACH CLASS

Class No.	Description	Members
Class 1	Caring for body parts problem	2
Class 2	Toileting problem	7
Class 3	Dressing problem	1
Class 4	Washing oneself, caring for body parts and dressing problem	12
Class 5	Washing oneself, caring for body parts, toileting and dressing problem	3
Class 6	Eating, drinking, washing oneself, caring for body parts, toileting, dressing, looking after one's health and looking after one's safety problem	29
Class 7	No problem	16

There are seven codes are considered for the participation component and impairment extent of activities are shown in table V.

TABLE V. IMPAIRMENT VALUES

Code	Value	Description
XXX.0	0-4%	NO impairment
XXX.1	5-24%	MILD impairment
XXX.2	25-49%	MODERATE impairment
XXX.3	50-95%	SEVERE impairment
XXX.4	96-100%	COMPLETE impairment
XXX.8		not specified
XXX.9		not applicable
XXX.9		not applicable

B. Data Preprocessing

After collecting the data, the implementation was done in R Programming version 4.1.2 using RStudio editor. It contains many in built packages. Data preprocessing is nothing but checking the dataset before going to model creation in case any missing values, irrelevant values, outliers, duplicated values. This SCADI dataset does not contain any missing, duplicated values and outliers.

C. Feature Selection using Boruta Algorithm

Feature selection is playing very important role before creating any model if the dataset contains too many features. This leads to poor performance, time and memory consuming. To avoid this need to choose relevant features by omitting irrelevant features which are not really influence the result anyway. It is not only helpful to increase the performance of the model and also reduce the complexity of the model. Boruta is one of the feature selection algorithms which working based on random forest. Figure 2 describes the three coloured attributes in which red colours are rejected features, yellow colours are tentative features and green colours are confirmed features. Three blue colours are shadowmin, shadowmean and shadowmax. Boruta performed 499 iterations in 34.25225 seconds and confirmed important 48 attributes.

These selected attributes are d.5100.4, d.5101.2, d.5101.3, d.5101.4, d.5102.2, d.5102.3, d.5102.4, d.5200.3, d.5201.3, d.5201.4, d.5202.3, d.5202.4, d.5203.3, d.5203.4, d.5204.3, d.5204.4, d.5205.3, d.5205.4, d.53000.0, d.53000.1, d.53000.4, d.53001.2, d.53001.4, d.53010.0, d.53010.1, d.53010.4, d.53011.2, d.53011.4, d.5400.2, d.5400.3, d.5400.4, d.5401.3, d.5401.4, d.5402.2, d.5402.3, d.5402.4, d.5403.2, d.5403.4, d.5404.0, d.5404.1,

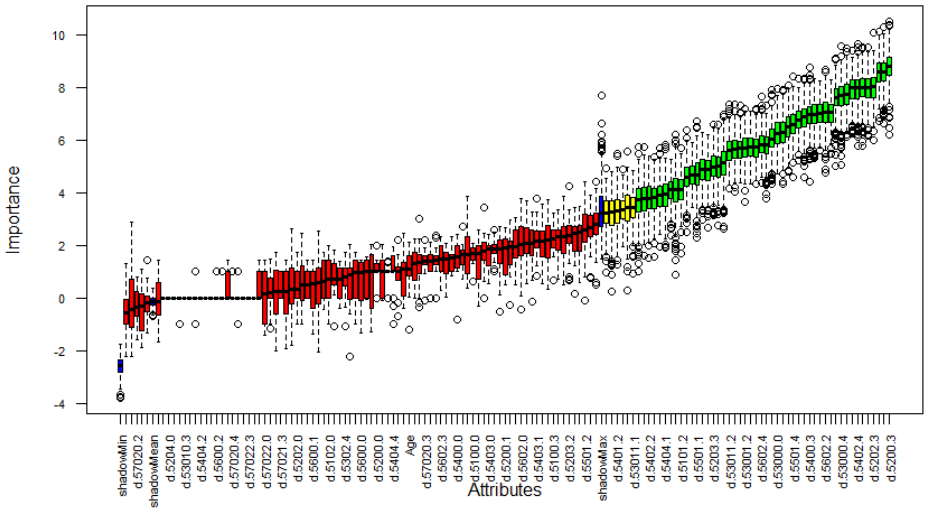

Figure 2. Plot for Confirmed, Tentative and Rejected attributes

d.5501.1, d.5501.3, d.5501.4, d.5602.1, d.5602.2, d.5602.4, d.5700.1, d.571.1. Table VI clearly explain the few records with values of meanImp, medianImp, minImp, maxImp, normHits and decision taken for tentative and confirmed features. Basically based on these values the decision has been taken by Boruta algorithm. Figure 3 shows the plot of rejected, confirmed, tentative features, shadowmin, shadowmean and shadowmax for different classifier run with their importance values.

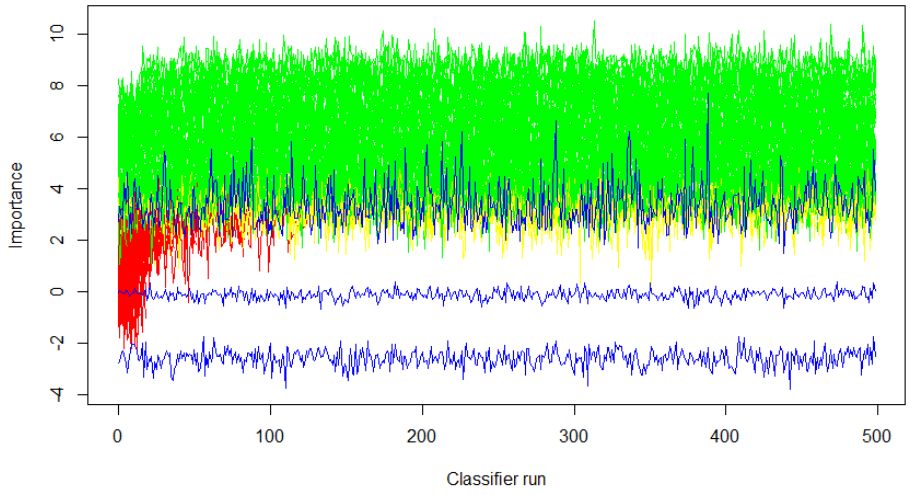

Figure 3. Plot of Confirmed, Tentative and Rejected attributes for different classifier run

Table VI clearly explain the values of meanImp, medianImp, minImp, maxImp, normHits and decision taken for tentative and confirmed features. Basically based on these values the decision has been taken by Boruta algorithm. Decision is based on normHits value, if it is greater than 0.6, confirmed; otherwise tentative features.

D. Model Creation

After selected confirmed and tentative features using Boruta algorithm, the next step is model creation. The packages e1071, ggplot2 and caret were imported. The dataset was divided into two sets. 80% of data considered as training dataset which was used to create the model and 20% of testing dataset used for checking the model. The model is created using SVM classifier in R programming. The svm() method was used to create the model and predict() method was used to predict the model. The SVM classifier with four kernels namely linear, radial, polynomial and sigmoid were created with whole dataset, confirmed selected (48) and confirmed & tentatively selected (54) features.

TABLE VI. SAMPLE DECISION OF FEATURES

Features	meanImp	medianImp	minImp	maxImp	normHits	Decision
d.5100.2	3.388132	3.40955	0.259296	5.552281	0.553106	Tentative
d.5200.3	8.757243	8.787478	6.203275	10.50437	1	Confirmed
d.5200.4	3.344867	3.336902	1.952495	4.872333	0.53507	Tentative
d.5201.3	7.962146	7.982149	6.137609	9.501267	1	Confirmed
d.53000.4	7.628961	7.672417	5.038185	9.574681	0.997996	Confirmed
d.53001.1	3.238793	3.233267	0.977531	4.938522	0.527054	Tentative
d.53001.2	5.689437	5.685118	3.620014	7.161829	0.967936	Confirmed
d.53001.4	8.560397	8.589259	6.628726	10.27316	1	Confirmed
d.53010.4	7.598254	7.615981	5.022668	9.072321	0.997996	Confirmed
d.53011.1	3.404674	3.436215	0.9347	4.838585	0.56513	Tentative
d.53011.2	5.589118	5.617954	4.160463	7.351957	0.963928	Confirmed
d.53011.4	8.576517	8.571363	6.709818	10.12678	1	Confirmed
d.5400.4	7.904686	7.959437	6.225494	9.373461	1	Confirmed
d.5401.2	3.268793	3.299912	1.242022	5.449844	0.51503	Tentative
d.5401.3	6.840152	6.860248	4.597167	8.338962	0.997996	Confirmed
d.5401.4	5.692195	5.705139	4.313565	7.074841	0.983968	Confirmed
d.5700.1	5.113717	5.117049	2.636118	6.940308	0.931864	Confirmed
d.5701.1	3.182597	3.250334	0.253117	4.801091	0.480962	Tentative
d.571.1	4.059187	4.091116	0.890023	6.202809	0.751503	Confirmed

E. Model Evaluation

The accuracy of the model was predicted in three different ways. Table VII depicts the performance of the model with four kernels with whole dataset and selected attributes. The SVM classifier outperformed with confirmed features. The linear kernel produced 97.14% with best accuracy.

TABLE VII. SVM PERFORMANCE FOR WHOLE DATASET AND CONFIRMED ATTRIBUTES WITH DIFFERENT KERNEL TYPE

SVM performance for whole dataset with different kernel type				
Kernel Type ->	Radial	Linear	Polynomial	Sigmoid
Number of Support Vectors	56	54	52	58
Accuracy	81.82%	81.82%	72.73%	54.55%
Kappa	0.7381	0.7443	0.6163	0.1912
Performance for confirmed 48 features with different kernel type				
Number of Support Vectors	38	35	43	40
Accuracy	91.43%	97.14%	94.29%	87.14%
Kappa	0.8815	0.961	0.9212	0.8213

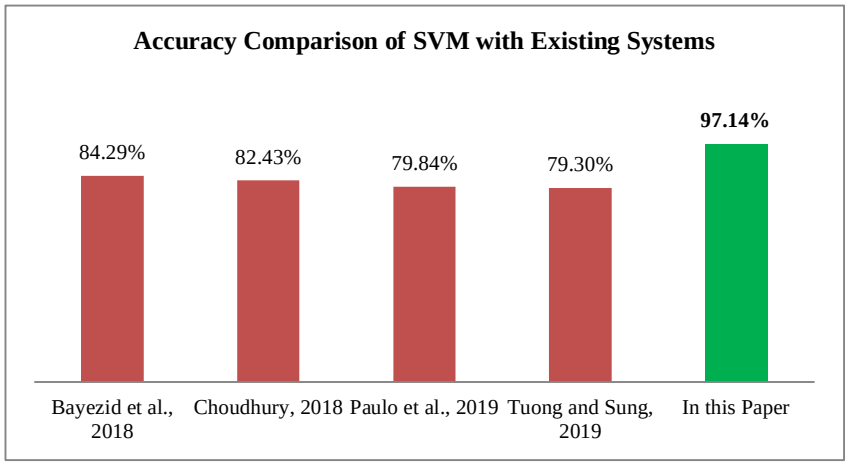

Figure 4. Comparison result of SVM models

Figure 4 shows the accuracy comparison of SVM classifiers with the existing work in which SVM classifier used to create model. Proposed system produced highest accuracy.

IV. CONCLUSIONS

An expert model was created using Support Vector Machine classifier to predict the self-care problems of children and youth with physical and moor disability in earlier stage. The well-known and only available SCADI (Self-Care Activities Dataset based on ICF-CY) dataset was utilized. Dimensionality reduction was used to select relevant features which are really needed for the models. The Boruta algorithm was used for feature selection. The Support Vector Machine model was developed using whole dataset and also 48 selected features. The model outperformed with **97.14%** on selected attributes rather than all attributes. The future enhancement is needed to find the solution to avoid the unbalanced problem of SCADI dataset.

REFERENCES

[1] Zarchi, Mohsen Sardari, SMM Fatemi Bushehri, and M. Dehghanizadeh. "SCADI: A standard dataset for self-care problems classification of children with physical and motor disability." International journal of medical informatics 114 (2018): 81-87.

[2] Islam, Bayezid, NI Md Ashafuddula, and Firoz Mahmud. "A machine learning approach to detect self-care problems of children with physical and motor disability." 2018 21st International Conference of Computer and Information Technology (ICCIT). IEEE, 2018.

[3] Keleş, Mümine Kaya, and ÜmitKılıç. "Artificial bee colony algorithm for feature selection on scadi dataset." 2018 3rd International Conference on Computer Science and Engineering (UBMK). IEEE, 2018.

[4] Chawathe, Sudarshan S. "Classifying Self-Care Activities of Children and Youths with Disabilities." 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2018.

[5] Liu, Lei, et al. "Feature selection based on feature curve of subclass problem." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[6] Souza, Paulo VC, et al. "Using hybrid systems in the construction of expert systems in the identification of cognitive and motor problems in children and young people." 2019 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE, 2019.

[7] Choudhury, Avishek. "Classification of functioning, disability, and health: ICF-CY self care (SCADI dataset) using predictive analytics.", In: Proceedings of the 2019 IISE Annual Conference. Edited by HE Romeijn AS, R. Thomas. Orlando: IISE. 2019.

[8] Bushehri, SMM Fatemi, and Mohsen SardariZarchi. "An expert model for self-care problems classification using probabilistic neural network and feature selection approach." Applied Soft Computing 82 (2019): 105545.

[9] Le, Tuong, and Sung WookBaik. "A robust framework for self-care problem identification for children with disability." Symmetry 11.1 (2019): 89.

[10] Omer Faruk Ertugrul, Necmettin Sezgin, "Classifying Self-Care Problems of Children with Physical and Motor Disability by Extreme Learning Machines", 2. ULUSLARARASI ZEUGMA BİLİMSEL ARAŞTIRMALAR KONGRESİ, 2019.

[11] Akyol, Kemal. "Comparing of deep neural networks and extreme learning machines based on growing and pruning approach." Expert Systems with Applications 140 (2020): 112875.

[12] Putatunda, Sayan. "Care2Vec: a hybrid autoencoder-based approach for the classification of self-care problems in physically disabled children." Neural Computing and Applications 32.23 (2020): 17669-17680.

[13] Syafrudin, Muhammad, et al. "A Self-Care Prediction Model for Children with Disability Based on Genetic Algorithm and Extreme Gradient Boosting." Mathematics 8.9 (2020): 1590.

[14] Syafrudin, Muhammad, et al. "Improving Efficiency of Self-care Classification Using PCA and Decision Tree Algorithm." 2020 International Conference on Decision Aid Sciences and Application (DASA). IEEE, 2020.

[15] Dardzińska-Głębocka, Agnieszka, and MałgorzataZdrodowska. "Analysis children with disabilities self-care problems based on selected data mining techniques." Procedia Computer Science 192 (2021): 2854-2862.

[16] Mweshi, Margaret M. "Use of the International Classification of Functioning, Disability and Health–Children and Youth (ICF-CY) in the Management of Children with Disabilities." IJNPT 2.1 (2016): 5-11.