ANALYSIS OF METHODS TO HANDLE MEDICAL SENSOR DATA TOWARDS HEALTH DISORDER IDENTIFICATION

S. Meruja and E.Y.A. Charles
Department of Computer Science, Faculty of Science, University of Jaffna
merujajoseph@gmail.com, charles.ey@univ.jfn.ac.lk

Introduction

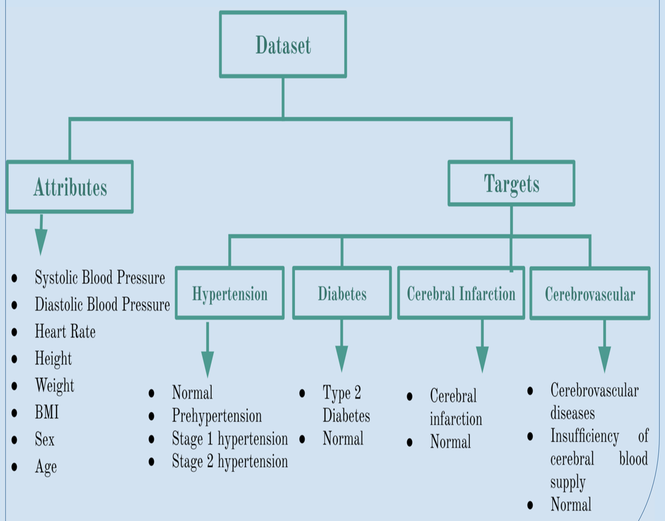
Medical sensors are important for monitoring the health condition of a person. The huge amount of data generated by these sensors, has a great potential for early detection of disorders. This study focuses on analysing methods to handle medical sensor data, particularly the data from Photo-plethysmograph (PPG). PPG quantifies the volumetric change of the heart by measuring the light transmission or reflection on arteries.

Objective

- Analyse various pre-processing approaches to handle issues arising in medical datasets, especially the imbalance of data items.
- Compare various classifiers to identify an optimal classifier for medical sensor data.

Dataset

- This study has used PPG-BP dataset containing features namely, Systolic Blood Pressure, Diastolic Blood Pressure, Heart Rate, Age, Sex and Body Mass Index (BMI) which were collected from 219 persons.
- Systolic blood pressure, diastolic blood pressure and heart rate were calculated from the PPG recordings.
- The dataset contains data for normal persons and persons with four disorders namely; hypertension, diabetes, cerebral infarction, and cerebrovascular disease.



Methodology

➔ Analyse the data and identify appropriate methods to represent data suitable for Machine Learning.

➔ Identify appropriate methods to handle class imbalances in the data set.

The dataset selected for this study is highly imbalanced. Dataset contains a higher number of records from normal persons. To avoid the effect of ***overfitting*** during learning, this study used:

Undersampling technique: Taking a random number of data with target value 'Normal' according to the ratio of other classes.

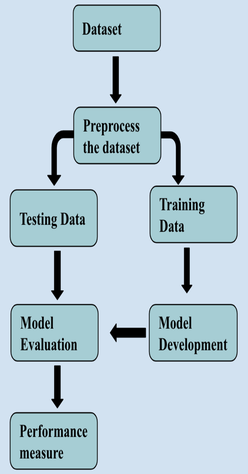
Duplicating the dataset: Make a copy of dataset and scaled it and added to the dataset. Again it leads to overfitting.

Undersampling and duplicating: Taking a random number of data with target value 'Normal' according to the ratio of other classes and a copy of the dataset is scaled and added to the dataset.

➔ Identify appropriate Machine learning methods suitable for the selected dataset. This study selected SVM, K-Nearest Neighbor and Naïve Bayes algorithms.

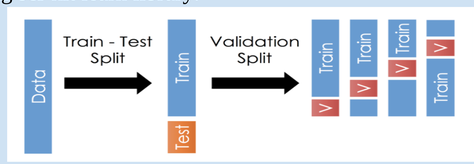
➔ Construction of classifiers - Relationship between attributes and target values. Four different models were constructed, each for one disorder.

➔ Evaluate the constructed models and revise the whole process to identify methods and models for optimal performance.



Experimental setup

Proposed methods and models were implemented and tested using sci-kit learn library.



- The data set is divided into four parts, each for identifying a particular disorder.
- As the standard practice, 70% of the dataset was used to train the models and 30% was used for testing.
- Feature Selection : Recursive Feature Elimination - To reduce the number of parameters used to construct the classifiers.
- Classifiers: Linear SVM-OvA, Linear SVM-OvO, K-Nearest Neighbour, and Naive Bayes

Testing Result

Accuracy, Precision, F1 score and Recall are used as the performance metric to validate the models and confusion matrix to analyse the performance of the proposed approaches.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$
$$\text{Precision} = TP / (TP + FP)$$
$$\text{Recall} = TP / (TP + FN)$$
$$\text{F1 Score} = 2*[(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})]$$

TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative

Table 1: The classification rate of the proposed method for each disease.

Classifier	*Hypertension*		*Diabetes*		*Cerebrovascular*		*Cerebral Infarction*	
	Overall	*N-Fold*	*Overall*	*N-Fold*	*Overall*	*N-Fold*	*Overall*	*N-Fold*
SVM-OvA	85.45%	86.79%	63.16%	59.89%	66.67%	**88.57%**	90.00%	87.50%
SVM-OvO	**98.18%**	**96.82%**	63.16%	59.89%	71.43%	74.29%	90.00%	80.00%
K-Nearest Neighbour	83.64%	83.22%	**84.21%**	57.81%	**90.48%**	62%	**100.00%**	83%
Naive Bayes	92.73%	90.91%	52.63%	**65.89%**	71.43%	85.71%	90.00%	**92.50%**

References

[1]. N.D.K.G. Dharmasiri and S. Vasanthapriyan, "Approach to Heart Diseases Diagnosis and Monitoring through Machine Learning and iOS Mobile Application," Proceedings of the International Conference on Advances in ICT for Emerging Regions ICTers, vol. 18, pp. 407–412, 2018.

[2]. Liang, Y. et al, "A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China," Sci. Data 5:180020 doi: 10.1038/sdata.2018..

[3]. Arrozag Ave, Hamdan Fauzan S. Rhandy Adhitya and Hasballah Zakaria, "Earlier detection of cardiovascular diseases with photoplethysmogram (PPG) sensor," Proceedings of the International Conference on Electrical Engineering and Informatics, pp. 676–681, 2015.

Discussion

➔ This study concentrated on two aspects:
  - ◆ Identifying the morphological information of PPG signals to identify disorders.
  - ◆ Implementation of machine learning models to identify the type of disorder.

➔ KNN provides the highest accuracy for identifying Diabetes (84.21%), Cerebral infarction (100%), and Cerebrovascular disease (90.48%) than other models for different K.

➔ Even though SVM is considered as a superior classifier compared to KNN and Naive Bayes it has achieved the highest accuracy only for identifying Hypertension (98.41%)

Conclusion

- This study analyzed the various methods and models on using PPG signals for Cardiovascular disorder identification.
- Learning models are able to perform better on imbalanced data sets in terms of prediction accuracy when the proposed techniques are applied.
- This study shows that the PPG signal data can be used for prediction of certain types of Cardiovascular disorders with adequate accuracy.