Dimensionality Reduction in Supervised Models-based for Heart Failure Prediction

Anna Karen Garate Escamilla[1], Amir Hajjam El Hassani[1] and Emmanuel Andres[2,3]
[1]*Nanomedicine Lab, Univ. Bourgogne Franche-Comte, UTBM, F-90010 Belfort, France*
[2]*Service de Médecine Interne, Diabète et Maladies métaboliques de la Clinique Médicale B, CHRU de Strasbourg, Strasbourg, France*
[3]*Centre de Recherche Pédagogique en Sciences de la Santé, Faculté de Médecine de Strasbourg, Université de Strasbourg (UdS), Strasbourg, France*

Keywords: Machine Learning, Heart Failure, Apache Spark, Feature Selection, PCA.

Abstract: Cardiovascular diseases are the leading cause of death worldwide. Therefore, the use of computer science, especially machine learning, arrives as a solution to assist the practitioners. The literature presents different machine learning models that provide recommendations and alerts in case of anomalies, such as the case of heart failure. This work used dimensionality reduction techniques to improve the prediction of whether a patient has heart failure through the validation of classifiers. The information used for the analysis was extracted from the UCI Machine Learning Repository with data sets containing 13 features and a binary categorical feature. Of the 13 features, top six features were ranked by Chi-square feature selector and then a PCA analysis was performed. The selected features were applied to the seven classification models for validation. The best performance was presented by the ChiSqSelector and PCA models.

1 INTRODUCTION

The WHO (World Health Organization, 2018) lists cardiovascular diseases as the leading cause of death worldwide with 17.7 million people dying every year. Heart diseases are affected by alcohol consumption, tobacco use, lack of exercise, an unhealthy diet and are present in people with high blood pressure, high blood glucose, overweight and obesity. A well-known cardiovascular disease is heart failure. Heart failure (HF) is a chronic condition present when the heart cannot pump enough blood to meet the necessity of the body. The American Heart Association lists the symptoms of HF such as shortness of breath, weight gain (1 or 2 kg. per day), fatigue, trouble sleeping, swelling in the legs, chronic cough and high heart rate (Heart, 2018). The diagnosis of heart failure can be a problem for the practitioners given its nature of being common or confused with the signs of aging.

The growth in the collection of medical data presents a new opportunity for doctors to improve the diagnosis of patients. In recent years, machine learning has become an important solution in the healthcare industry. Machine learning is an analytical tool that works to help users identify patterns and relationships by learning from experience. It is used when the task is very large and complex to program, such as the transformation of medical files into knowledge, pandemic predictions and genomic data analysis (Shalev-Shwartz et al., 2016).

In the past, different studies have been done using machine learning techniques to diagnose different cardiac issues and predict the outcome. The study of Rahhal et al. (2016) proposed a classification of electrocardiogram (ECG) signals through a deep neural network (DNN). Khalaf et al. (2015) classified cardiac arrhythmias using computer-aided diagnostic (CAD) systems to categorize five types of beats. The prediction was with support vector machine (SVM), obtaining an accuracy of 98.60% with raw data, 96.30% with PCA and 97.60% with Fisher Score (FS). Guidi et al. (2014) proposed a clinical decision support system (CDSS) for the analysis of HF. The best accuracy was 87.6% using the CART model. Parthiban and Srivatsa (2012) used a SVM technique to diagnose heart disease in patients with diabetes, obtaining an accuracy of 94.60%.

Escamilla, A., El Hassani, A. and Andres, E.
Dimensionality Reduction in Supervised Models-based for Heart Failure Prediction.
DOI: 10.5220/0007313703880395
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 388-395
ISBN: 978-989-758-351-3

The main problem of machine learning is the high dimensionality (Domingos, 2012). The key to the success of machine learning models is to select the best features. It can be observed in the literature that the use of feature selection techniques helped the performance of a classification algorithm in the prediction of HF. Dun et al. (2016) used deep learning, random forest, logistic regression, SVM and neural network with hyperparameters and feature selection to predict the presence of heart disease, obtaining an accuracy of 78%. Yaghouby et al. (2009) classified arrhythmias using the generalized discriminant analysis (GDA) and the multilayer perception (MLP) neural network with an accuracy of 100%. Rajagopal et al. (2017) presented a classification of cardiac arrhythmia using five different linear and non-linear unsupervised dimensionality reduction techniques combined with a probabilistic neural network (PNN) classifier. The PNN classifier and the fast independent component analysis (fastICA) obtained the best result with 99.83%. Singh et al. (2018) computed better results in the detection of coronary heart disease using reduction functions with 100% accuracy. Asl et al. (2008) presented a classification that used 15 features extracted from heart rate variability (HRV) signal. The authors reduced the features to five using a GDA technique and increased the accuracy to 100% when combined with the SVM classifier.

This paper proposes the combination of classification models with dimensionality reduction techniques to achieve two main objectives: (1) to learn the best feature representation of the data set used; and (2) to use machine learning techniques as a classifier to obtain the best possible prediction. The data set used to achieve this purpose came from the UCI Machine Learning Repository (UCI, 2018) and is computed with seven classifiers: logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perception, one-vs-rest and Naïve Bayes. The feature selection technique of Chi-square is implemented and used by PCA.

Remaining of this paper is organized as follows. A short summary of the models used are explained in Section 2. Detail descriptions of the methodology are presented in Section 3. Experimental results are reported in Section 4. Finally, Section 5 concludes the work.

2 THEORICAL BACKGROUND

The classifiers models used in this paper are presented in this section.

2.1 Logistic Regression

Logistic regression (Hastie et al., 2017) is a binary classification response used to describe information and explain the relationship between dependent and independent variables. For a binary classification, the model makes predictions by applying the logistic function

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

2.2 Decision Tree

Decision tree is a classification and regression method commonly used for machine learning because its nature of being easy to interpret. The tree predicts the label for each partition (leaf), and each one is chosen by selecting the best split of the different possible splits. Each tree node is chosen from the set $argmax IG_s(D, s)$ where $IG(D, s)$ is the information obtained when a split s is applied to a dataset D (Marsland et al., 2015).

2.3 Random Forest

Random forest (Breiman, 2001) is a collection of decision trees predictors in which each tree depends on the value of an independent random vector. The training algorithm works in a parallel and random mode, making each decision tree different and with a reduction in variance.

2.4 Gradient-Boosted Tree

Gradient-Boosted Trees (GBTs) are ensembles of decision trees which minimize a loss function (Friedman, 1999). The mechanism used to reduce the loss function in the training data is given by

$$f(\mathrm{x}) = 2\sum_{i=1}^{N} log\left(1 + exp\left(-2y_i F(x_i)\right)\right) \tag{2}$$

where N=number of instances, y_i=label of instance i, x_i=features of instance i, $F(x_i)$=model's predicted label for instance i.

2.5 Multilayer Percepton

Multilayer perceptron classifier (Hornik, 1991) is based on the feedforward artificial neural network and consists of multiple layers of nodes. Each layer is connected to the next layer. The input

data is represented by the nodes in the input layer. The other nodes map the input to the output by combining the weight w and the bias b of the node. This is written as a matrix with $K+1$ layers as

$$y(x) = f_k(\dots f_2(w_2^T f_1(w_1^T x + b_1) + b_2) \dots + b_k) \quad (3)$$

The nodes in intermediate layers use a sigmoid function given by

$$f(z_i) = \frac{1}{1 + e^{-z_i}} \quad (4)$$

The nodes in the output layer use softmax function given by

$$f(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} \quad (5)$$

where N corresponds to the number of classes.

2.6 One-vs-Rest

One-vs-Rest is a classifier that creates a binary classification problem for each class. One-vs-Rest converts one class as positive and the rest of the classes as negative. The classifier with the highest value will be the output.

2.7 Naïve Bayes

Naïve Bayes is a classifier based on the theorem of Bayes with strong independence assumptions between the features. It works with the assumption of using observation of the problem to make a prediction (Marsland, 2018).

3 METHODOLOGY

The data set used in the research is the "Heart Disease Data set" from the UCI Machine Learning Repository. The data set contains 76 features, but most of the existing articles used only the subset of 14 features described in Table 1. The categorical feature Num contains whether a patient has a presence or absence of a heart disease. The categorical features 1, 2, 3 and 4 of the original data set were transformed in one that is the presence (1) of heart disease.

The data sets used are from hospitals in Cleveland, Hungarian, Switzerland and Long Beach VA. This study adds one more data set: Cleveland-Hungarian (a combination of Cleveland and Hungarian data sets with 597 patients). The most common data set in other studies is Cleveland, which has great data quality. On the contrary, Hungarian, Switzerland and Long Beach VA have

Table 1: Features of Heart Disease Data set.

Number	Code	Feature	Description
1	Age	Age	Age in years
2	Sex	Sex	1=male; 0=female
3	Cp	Chest pain type	1= typical angina; 2=atypical angina; 3=non-angina pain; 4= asymptomatic
4	Trestbps	Resting blood pressure (mg)	At the time of admission in hospital
5	Chol	Serum cholesterol (mg)	
6	Fbs	Fasting blood sugar>120 mg/dl	1=yes; 0=no
7	Restecg	Resting electrocardiographic results	0=normal; 1= ST-T wave abnormal; 2=left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	
9	Exang	Exercise induced angina	1=yes; 0=no
10	Oldpeak	ST depression induced by exercise relative to rest	
11	Slope	The slope of the peak exercise ST segment	1=upsloping; 2=flat; 3=downsloping
12	Ca	Number of major vessels (0-3) colored by fluoroscopy	
13	Thal	Exercise thallium scintigraphy	3=normal; 6=fixed defect; 7=reversible defect
14	Num	The predicted attribute	0=no presence; 1=presence

several missing values and are less used. The Table 2 contains the number of patients per data set and the quality of their data. The considerations taken for data cleansing were to assign a unique category for the missing values and create rules considering the coherence of the data. An example of this is that a patient cannot have a cholesterol equal to zero.

The models and algorithms used were computed using the MLlib guide of Apache Spark 2.2.0 and programmed in Java language (Spark, 2018). This library provides the classification and dimensionality reduction tools to obtain the performance.

Table 2: Heart Disease Data set.

Data Set	Patients	Quality of the data
Cleveland	303	Complete
Hungarian	294	Some feature are incomplete (slope, ca and thal)
Switzerland	123	Some feature are incomplete (thalach, chol, exang, slope, ca and thal)
Long Beach VA	200	Some feature are incomplete (fbs, ca and thal)

3.1 Dimensionality Reduction

Dimensionality reduction (Domingos, 2018) is the process of reducing the number of variables under consideration. It can be used to extract latent features of raw data sets or compressing data while maintaining the structure. This research proposed two different dimensionality reduction methods, for the feature selection, the Chi-square test of independence was selected and for feature extraction, the principal component analysis (PCA). After several attempts, Chi-square test, with $k=6$, performs better than other feature selection techniques in the literature. The Chi-square test order features based on the class and filters the top features of which the class label depends on the most. ChiSqSelector (ChiSq) of Apache Spark MLlib is used for feature selection in model construction.

The list of the reduced set of features is shown in Table 3. The order of the features in each row is selected from most to least important. Further, the six features were validated using seven classifiers depicted in the next section. It can be observed that chest pain (cp) is common in all data sets with the exception of Switzerland. Similar to cp, cholesterol (chol), maximum heart rate achieved (thalach), the ST depression induced by exercise relative to rest (oldpeak) are common. The values of exercise induce angina (exang), exercise thallium scintigraphy (thal) and the number of major vessels colored by fluoroscopy (ca) were the results of the non-invasive test to determine if the patient has heart failure. Due to the poor quality of data, with the exception of Cleveland, these features do not have high rank.

PCA is a statistical procedure that converts the correlated features into a new set of uncorrelated features with the aim of losing the less amount of information. The PCA class trains a model to project vectors to a low-dimensional space. After several test, the best result for PCA was given by the features selected of ChiSq to create the components instead of using the raw data.

Table 3: Features selected by ChiSq.

Data set	List of features
Cleveland	chol, thalach, oldpeak, thal, cp, ca
Hungarian	chol, slope, exang, oldpeak, thalach, cp
Long Beach VA	chol, thalach, age, trestbps, oldpeak, cp
Switzerland	thalach, oldpeak, age, cp, trestbps, restecg
Cleveland-Hungarian	chol, oldpeak, cp, exang, slope, thalach

3.2 Classifiers Proposed

For this research, the libraries of ML Spark are used to make the predictions. The classification models were computed with default value of their hyperparameters. The models are: (1) decision tree (DT); (2) gradient-boosted tree (GBT); (3) logistic regression (LOG); (4) multilayer percepton (MPC); (5) Naive Bayes (NB); (6) one-vs-rest (OvR); and (7) random forest (RF).

For experimentation, the data sets are divided in 70% for training, of which 80% is used for training and 20% for validation, and 30% for testing. The classifiers run 10 times and evaluate the accuracy, precision, recall and F1 score of the categorical feature, observing the percentage of the correct classification. The confusion matrix reports the basic terms used by these evaluations: (1) true positives (TP) are cases in which the patients have heart disease and are correctly predicted; (2) true negatives (TN) are patients who do not have a heart disease and are predicted as negative; (3) false positives (FP) are patients predicted as positive, but do not have heart disease; and (4) false negatives (FN) are patients predicted as negative, but they have a heart disease.

Table 4: Cleveland predictions.

Features	Performance	DT	GBT	LOG	MPC	NB	OvR	RF
Raw data	Accuracy (%)	84.3	83.5	88.7	84.5	72.7	89.7	90.4
	Precision (%)	88.6	86.8	88.5	78.4	67.9	94.1	83.8
	Recall (%)	77.5	78.6	85.2	85.1	86.4	84.2	88.6
	F1 (%)	82.7	82.5	86.8	81.6	76.0	88.9	86.1
ChiSq-PCA	Accuracy (%)	84.3	83.8	89.3	87.5	78.0	90.9	87.4
	Precision (%)	86.8	78.9	92.3	89.5	90.0	94.6	83.3
	Recall (%)	78.6	85.7	85.7	82.9	64.3	88.1	89.7
	F1 (%)	82.5	82.2	88.9	86.1	75.0	91.2	86.4
ChiSq	Accuracy (%)	85.1	83.9	88.6	90.3	75.3	88.4	87.8
	Precision (%)	86.1	85.0	93.1	89.3	71.9	89.7	88.4
	Recall (%)	77.5	79.1	81.8	86.2	83.7	83.3	79.2
	F1 (%)	81.6	81.9	87.1	87.7	77.4	86.4	83.5

Table 5: Hungarian predictions.

Features	Performance	DT	GBT	LOG	MPC	NB	OvR	RF
Raw data	Accuracy (%)	83.5	86.0	92.0	87.8	85.9	89.7	90.5
	Precision (%)	76.5	83.8	95.8	86.7	78.8	81.1	88.9
	Recall (%)	81.3	79.5	76.7	72.2	81.3	90.9	80.0
	F1 (%)	78.8	81.6	85.2	78.8	80.0	85.7	84.2
ChiSq-PCA	Accuracy (%)	85.9	86.7	91.0	86.3	83.8	89.0	86.6
	Precision (%)	75.9	75.0	95.5	80.8	70.9	86.2	89.3
	Recall (%)	88.0	87.5	75.0	77.8	84.6	80.6	71.4
	F1 (%)	81.5	80.8	84.0	79.2	83.3	84.4	80.0
ChiSq	Accuracy (%)	86.0	84.8	92.2	89.2	84.9	89.8	89.8
	Precision (%)	88.5	85.0	85.0	81.1	79.5	88.5	95.8
	Recall (%)	71.9	85.0	85.0	81.1	86.1	79.3	71.9
	F1 (%)	79.3	85.0	85.0	81.1	82.7	83.6	82.1

The accuracy rate is computed using the formula given by

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (6)$$

The precision is the positive predicted value defined by

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

The recall is defined as the proportion of patients with heart disease correctly identified given by

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

F1 score is given by the precision in Eq. (7) and recall in Eq. (8), considering an harmonic average defined by

$$F1\ score = 2\left(\frac{PrecisionxRecall}{Precision + Recall}\right) \quad (9)$$

4 RESULTS

The data sets are computed using the supervised machine learning algorithms. These 13 features are reduced to six new features or components using: (1) PCA algorithm with the features of ChiSq and (2) the top six features of the ChiSq. Finally, the classifiers validate the performance.

Table 4 contains the best performance of the Cleveland data set. In most cases, the best accuracy and F1 score improves when dimensionality reduction techniques are applied, with the exception of RF that computes the best accuracy using raw data with 90.4% accuracy and 86.1% F1 score. The distribution of information in Cleveland is uniform, which leads to similar accuracy and F1score. Overall, the best performance was using ChiSq-PCA-OvR with an accuracy of 90.9%, a precision of 88.1%, a recall of 88.1% and a F1 score of 91.2%. The most notable improvements presented by the dimensionality reduction techniques, compared to the raw data, were for PCA the computations of

Table 6: Long Beach VA predictions.

Features	Performance	DT	GBT	LOG	MPC	NB	OvR	RF
Raw data	Accuracy (%)	84.3	80.0	85.9	83.1	76.7	82.6	84.2
	Precision (%)	88.9	80.0	80.0	66.7	43.8	57.1	50.0
	Recall (%)	53.3	28.6	50.0	42.9	50.0	30.8	27.3
	F1 (%)	66.7	42.1	61.5	52.2	46.7	40.0	35.3
ChiSq-PCA	Accuracy (%)	82.0	80.6	84.3	85.9	71.0	83.3	85.2
	Precision (%)	57.9	66.7	87.5	63.6	43.7	75.0	72.7
	Recall (%)	78.6	52.6	41.2	50.0	43.7	40.0	50.0
	F1 (%)	66.7	58.8	56.0	56.0	43.7	52.2	59.3
ChiSq	Accuracy (%)	86.0	79.4	85.1	83.6	76.3	83.3	87.5
	Precision (%)	77.8	50.0	80.0	75.0	52.9	58.3	85.7
	Recall (%)	58.3	57.1	40.0	42.9	60.0	58.3	54.5
	F1 (%)	66.7	53.3	53.3	54.5	56.3	58.3	66.7

Table 7: Switzerland predictions.

Features	Performance	DT	GBT	LOG	MPC	NB	OvR	RF
Raw data	Accuracy (%)	94.1	97.0	96.8	96.3	73.2	97.1	97.1
	Precision (%)	50.0	100.0	100.0	100.0	26.7	100.0	100.0
	Recall (%)	100.0	50.0	75.0	33.3	100.0	50.0	33.3
	F1 (%)	66.7	66.7	85.7	50.0	42.1	66.7	50.0
ChiSq-PCA	Accuracy (%)	94.3	93.8	100.0	97.3	97.0	97.3	97.4
	Precision (%)	66.7	50.0	100.0	100.0	100.0	100.0	100.0
	Recall (%)	66.7	50.0	100.0	33.3	75.0	33.3	33.3
	F1 (%)	66.7	50.0	100.0	50.0	85.7	50.0	50.0
ChiSq	Accuracy (%)	93.5	94.7	96.9	100.0	92.5	95.8	97.4
	Precision (%)	100.0	50.0	100.0	100.0	100.0	100.0	100.0
	Recall (%)	33.3	50.0	33.3	100.0	20.0	33.3	50.0
	F1 (%)	50.0	50.0	50.0	100.0	33.3	50.0	66.7

LOG and OvR, with an increase of 0.6% and 1.2% respectively. In the case of ChiSq were DT and MPC, with an increase of 0.8% and 5.8%.

The best results of the Hungarian data set are presented in Table 5. The best accuracy are computed using dimensionality reduction techniques in almost all the cases with the exception of RF and NB, which presented the best accuracy with 90.5% and 85.9% respectively. In general, the best accuracy results are computed by ChiSq-LOG with 92.2% accuracy and 85.0% of precision, recall and F1 score. The lack of uniform distribution shows that ChiSq-LOG has a better performance compared to raw data, which obtains a poor recall with 76.7%. Even if the performance is similar between raw data and dimensionality reduction techniques, ChiSq present the most remarkable results.

Table 6 shows the best results from the Long Beach VA data set. In general, dimensionality reduction techniques present better results than raw data, with the exception of LOG and NB that compute the best accuracy with 85.9% and 76.7% respectively. Overall, ChiSq-RF calculates the best accuracy with 87.5% and a recall and F1 score with one of the highest values with 85.7% and 66.7% respectively. In Long Beach VA, the results of precision, recall and F1 score were considerably low, this was due to a small rate of true positives and, therefore, this data set and models have a poor performance. The most notable improvements are, compared to the raw data, for PCA the computation of MPC with an increase of 2.8%. ChiSq increases 1.7% with DT and 3.3% with RF.

The best performance of the Switzerland data set is shown in Table 7. The Switzerland data set has better accuracy than the other data set presented, this can be explained by the presence of heart disease in 113 of the 123 patients, which means that the database is unbalanced. Due to the lack of uniformity, the tests obtain the greatest gap between accuracy and F1 score with a difference of around 40%. PCA compute the best accuracy in most of the cases, except for GBT with 97% using raw data. The best results are presented by ChiSq-PCA-LOG and ChiSq-MPC without errors, obtaining an accuracy, precision, recall and F1 score of 100%.

Table 8 presents the best results for the Cleveland-Hungarian data set. In general, ChiSq

Table 8: Cleveland-Hungarian predictions.

Features	Performance	DT	GBT	LOG	MPC	NB	OvR	RF
Raw data	Accuracy (%)	82.0	83.7	83.9	83.7	68.6	86.0	87.6
	Precision (%)	78.0	76.5	83.6	79.0	58.5	77.0	87.2
	Recall (%)	76.7	81.3	76.7	85.3	92.0	87.7	85.0
	F1 (%)	77.3	78.8	80.0	82.1	71.5	82.0	86.1
ChiSq-PCA	Accuracy (%)	82.1	81.4	87.0	84.4	66.9	85.7	79.6
	Precision (%)	77.5	81.2	86.7	77.8	58.9	82.1	82.7
	Recall (%)	77.5	70.9	79.3	86.3	91.2	82.1	69.7
	F1 (%)	77.5	75.7	82.8	81.8	71.6	82.1	75.6
ChiSq	Accuracy (%)	84.1	82.5	86.5	85.6	80.3	86.7	84.6
	Precision (%)	73.6	83.2	88.9	79.1	75.4	81.4	89.1
	Recall (%)	85.5	77.0	77.8	84.1	73.1	83.8	74.0
	F1 (%)	79.1	80.0	83.0	81.5	74.2	82.6	80.9

compute better results over raw data and PCA, except for GBT with 83.7% accuracy and RF with 87.6% accuracy. Overall, the best performance is RF using raw data with an accuracy of 87.6%, presenting remarkable values in precision, recall and F1 with 85%, 87.2% and 86.1% respectively.

The most outstanding comparison with all the features is the increase of PCA by 3.1% with LOG, ChiSq by 2.1% with DT, 1.9% with MPC, 11.7% with NB and 0.7 with OvR.

From the above results, almost all the classifiers work better using dimensionality reduction techniques. Only in one of the data sets, the best result is obtained using raw data with the RF compiler. In general, the dimensionality reduction techniques computed the best increment in accuracy using DT and LOG.

The results of the machine learning models combined with the dimensionality reduction techniques were: (1) the results of GBT and NB did not improve in most cases to generalize that the use of dimensionality reduction is better; (2) DT, LOG, MPC and OvR improved when it was used with PCA and ChiSq, LOG obtained better results with PCA and MPC with ChiSq; and (3) in most of the cases RF had the best results using all the features than a dimensionality reduction technique. In terms of quality of information, when the data is more complete, as in the case of Cleveland, the results of PCA are better than incomplete data sets.

The performance comparison of the Cleveland data set is given in Table 9. Based on this comparison, ChiSq-PCA-OvR approach had better accuracy than the literature methods. Comparing the same methods, the best accuracy in this research is given by decision tree with 85.1% using ChiSq, logistic regression with 89.3% using ChiSq-PCA, Naïve Bayes with 78.0% using ChiSq-PCA, random forest with 90.4% using raw data and multilayer percepton with 90.3% using ChiSq. With the exception of Naive Bayes, the results obtained show improvement compared to the proposed by the literature.

Table 9: Performance comparison of Cleveland.

Author	Method	Accuracy
Mutyala, et al. (2018)	Decision Tree C4.5	83.40%
Khanna, et al. (2015)	Logistic Regression	84.80%
Kodati, et al. (2018)	Naive Bayes	83.70%
Khan, et al. (2016)	Random Forest	89.25%
Ziasabounchi, et al. (2014)	ANFIS- Neuronal Network + Fuzzy rules in 5 layers	85.00%

5 CONCLUSIONS

In this paper, we proposed the use of dimensionality reduction techniques with machine learning classifiers to predict whether a patient has HF or not. The results presented by the ChiSq selector of Apache Spark were marvelous. The features that were persistent in all the data sets were chest pain (cp), cholesterol (chol), maximum heart rate achieved (thalach) and the ST depression induced by exercise relative to rest (oldpeak). These features most be considered important in the detection and analysis of HF. A disadvantage in the feature selection is the lack of data quality, especially in Switzerland and Long Beach VA.

PCA obtained better results with the features of ChiSq and when the data set does not have many null values. The experimental results obtained with

the classifiers improve, except with random forest that showed a better accuracy and F1 score when it was used with all the features. Overall, ChiSq and PCA obtained the highest accuracy, precision, recall and F1 score. LOG and RF were the classifiers that computed the best performance.

In general, the greatest problem with the models was the false negatives, this is important to consider, it is better to have a good classification of the false negatives than the false positives. For future development, some experimental work will attempt to model the physiological HF problem, which is difficult to do with few features. In addition, these models will be replicated in a big data health environment and test its functioning with massive databases.

REFERENCES

Asl, B., Setarehdan, S. and Mohebbi, M. 2008. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artificial Intelligence in Medicine*, 44(1), pp.51-64.

Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), p.78.

Dun B., Wang E., Majumder S. 2016. Heart Disease Diagnosis on Medical Data Using Ensemble Learning.

Friedman, J. H. 1999. *Greedy Function Approximation: A Gradient Boosting Machine*. Annals of Statistics. 29. 1189-1232. 10.2307/2699986.

Guidi, G., Pettenati, M., Melillo, P. and Iadanza, E. 2014. A Machine Learning System to Improve Heart Failure Patient Assistance. *IEEE Journal of Biomedical and Health Informatics*, 18(6), pp.1750-1756.

Hastie, T., Tibshirani, R. and Friedman, J. 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

HEART. 2018. Heart Failure. [online] Available at: http://www.heart.org/HEARTORG/Conditions/HeartFailure/HeartFailure_UCM_002019_SubHomePage.jsp [Accessed 16 Jun. 2018].

Khan, S.S. 2016. Prediction of Angiographic Disease Status using Rule Based Data Mining Techniques. *Biological Forum An International Journal*, 8(2), pp 103-107.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks,* 4(2):251–257.

Khanna, D., Sahu, R., Baths, V. and Deshpande, B. 2015. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. *International Journal of Machine Learning and Computing*, 5(5), pp.414-419.

Khalaf, A., Owis, M. and Yassine, I. 2015. A novel technique for cardiac arrhythmia classification using spectral correlation and support vector machines. *Expert Systems with Applications*, 42(21), pp.8361-8368.

Kodati, S. 2018. Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Global Journal of Computer Science and Technology*, 18(1).

Mutyala, N. 2018. Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools. *International Journal of Scientific Research in Computer Science*, Engineering and Information Technology, 3(3).

Marsland, S. 2015. *Machine Learning: an Algorithmic Perspective*. Chapman and Hall/CRC.

Parthiban, G. and K. Srivatsa, S. 2012. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. *International Journal of Applied Information Systems*, 3(7), pp.25-30.

Rahhal, M., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F. and Yager, R. 2016. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345, pp.340-354.

Rajagopal, R. and Ranganathan, V. 2017. Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification. *Biomedical Signal Processing and Control*, 34, pp.1-8.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5-32.

Shalev-Shwartz, S., Ben-David, S. 2016. Understanding machine learning: From theory to algorithms. *New York: Cambridge University Press*.

Singh, R., Saini, B. and Sunkaria, R. 2018. Detection of coronary artery disease by reduced features and extreme learning machine. *Clujul Medical*, 91(2), p.166.

Spark.apache.org. 2018. Linear Methods - RDD-based API - Spark 2.0.0 Documentation. [online] Available at: https://spark.apache.org/docs/2.0.0/mllib-linear-methods.html#logistic-regression [Accessed 20 Jun. 2018].

UCI. 2018. UCI Machine Learning Repository: Heart Disease Data Set. [online] Available at: http://archive.ics.uci.edu/ml/datasets/heart+disease [Accessed 20 Jun. 2018].

World Health Organization. 2018. Cardiovascular diseases (CVDs). [online] Available at: http://www.who.int/cardiovascular_diseases/en/ [Accessed 19 Jun. 2018].

Yaghouby, F., Ayatollah, A., Soleimani, R. 2009. Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal. *World Applied Sciences Journal*, 6 (11), pp.1547-1554.

Ziasabounch, N. 2014. ANFIS Based Classification Model for Heart Disease Prediction. *International Journal of Electrical & Computer Sciences*, 14 (2), pp.7-12.