International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Ascendable Clarification for Coronary Illness Prediction using Classification Mining and Feature Selection Performances

D. Haripriya[1], Dr. M. Lovelin Ponn Felciah[2]

[1]Research Scholar, [2]Assistant Professor

[1,2]Department of Computer Science, Bishop Heber College, Trichy, Tamil Nadu, India

How to cite this paper: D. Haripriya | Dr. M. Lovelin Ponn Felciah "Ascendable Clarification for Coronary Illness Prediction using Classification Mining and Feature Selection Performances" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1588-1592, **https://doi.org/10.31142/ijtsrd26690**

IJTSRD26690

ABSTRACT

Coronary disease is predicted by classification technique. The data mining tool WEKA has been exploited for implementing Naïve Bayes classifier. Proposed work is trapped with a specific end goal to enhance the execution of models. For improving the classification accuracy Naïve Bayes is combined with Bagging and Attribute Selection. Trial results demonstrated a critical change over in the current Naïve Bayes classifier. This approach enhances the classification accuracy and reduces computational time.

KEYWORDS: *Data mining, Heart diseases, WEKA, classification, Naïve Bayes, Bagging*

1. INTRODUCTION

The progress of data innovation, structure coordination and also programming headway, frameworks have shaped a creative period of multifaceted computer system. Data innovation authorities have been offered few difficulties by these systems. An instance of such structure is the health care services administrations system. As of late, there has been an expanded thoughtfulness regarding make use of the progress of Data mining propels in social insurance systems. Hence, the objective of the present exertion is to discover the parts of utilization of human services information for help of individuals by strategy for machine learning moreover Data mining methods. The main aim is to recommend a mechanized framework for diagnosing heart disease by considering prior data and information.

A noteworthy test going up against social insurance affiliations for example clinics, therapeutic centers are the acquirement of value administrations at sensible costs. Quality service suggests diagnosing patients precisely and managing medications that are increasingly successful. Poor clinical decision can provoke to poor results which are in this manner unacceptable. Human services associations can decrease costs by achievement of computer based information or potentially choice emotionally supportive networks. Social insurance administrations information is immense as it consolidates patient records, asset the board data and refreshed data. Human service affiliations must have ability to separate data. Treatment records of numerous patients can be put away in modernized manner; besides information mining strategies may help in discovering a couple of crucial and fundamental request related with human services associations. Data mining is a significant advance of KDD for example information disclosure from database. KDD involves an iterative succession of data cleaning, data integration, data decision, data mining design acknowledgment besides information introduction. In points of interest, data mining may be accomplished utilizing classification, clustering, prediction, association and time Series analysis. The Fig1 [1] shows that the normal artery and the diseased artery.

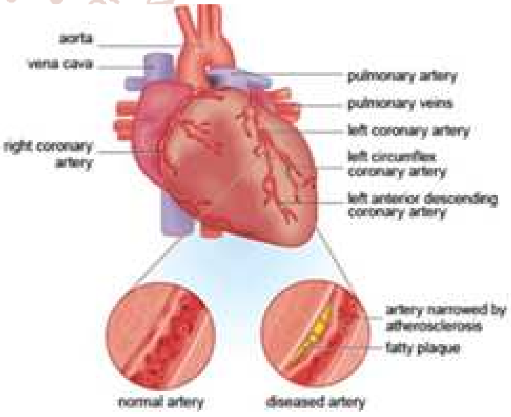


Fig 1 Heart Arteries

1.1. Heart Diseases:

Heart is important part or an organ of the body. Life is responsible to capable working of heart. In the occasion that activity of heart isn't appropriate, it will impact the other body parts of human, for example, mind, kidney, etc. Heart is basically a pump, which pumps the blood through the body. If in the event that blood in body is insufficient, at that point many organs like cerebrum suffer and if heart stops working by, death occurs inside minutes. Life is absolutely subject to effective working of the heart. The term Heart sickness implies sickness of heart and vessel structure inside it.

There are many elements which form the danger of Heart infection:

- Family history of coronary illness
- Smoking
- Poor eating methodology
- Obesity
- Physical inertia
- High pulse
- Cholesterol

2. High blood Cholesterol Literature survey

H. Benjamin Fredrick et al. [2] in this paper the general goal of the work is to anticipate all the more precisely the event of coronary illness utilizing data mining systems. In this research work, the UCI data repository is utilized for playing out the similar analysis of three algorithms, for example, Random Forest, Decision trees and Naive Bayes. From the research work, it has been tentatively demonstrated that Random Forest gives ideal outcomes as contrast with Decision tree and Naive Bayes. The Future work of this exploration work can be had to deliver an effect in the exactness of the Decision Tree and Bayesian Classification for extra improvement in the wake of applying genetic algorithm so as to decrease the real data for procuring the ideal subset of property that is sufficient for coronary illness expectation. The automation of coronary illness forecast utilizing genuine real time data from health care associations and organizations which can be manufactured utilizing big data. They can be streaming data and by utilizing the data, investigation of the patients progressively can be prepared.

Kanika Pahwa et al. [3] proposed this paper the fundamental intention of this research is to classify the data in two classes either in positive or in negative outcome for coronary illness. A hybrid approach of feature selection is embraced to enhance the classification problem, consolidated consequences of SVM-RFE and, gain-proportion are utilized to get subset of highlights and evacuate unessential or repetitive component. On subset of highlights naive bayes and random forest are connected to characterize them into presence or absence of disease. It has been appeared in results that precision improved for the two classifiers when connected to chosen selected features. Proposed approach of feature selection not just reduced size of dataset yet additionally upgraded the execution of both the classifiers models.

Shahed Anzarus Sabab et al. [4] designed this paper prediction of the coronary illness framework is created by joining Naïve Bayes and K-Means algorithm related to Hadoop. The develop software take out the learning from historical database made by medicinal specialist or the clinical consideration units. In initial step preparing of informational index is required then this informational collection is approved against already accessible preparing dataset for genuine working of the developed software. As created approach in hybrid it produces precision of accuracy of 85%.

3. DATA SOURCES

The input data has been collected from the UCI Repository. The data size is 4645 records and 11 attributes.

Table 1 shows that brief description of the data set have been considered.

Data set	No of attributes	No of Instances
UCI Repository	11	4645

Table 1: Dataset Description

The attribute description as shown in the table2:

S. No	Name	Description
1	Age	Age in years
2	Sex	Male or Female
3	BP	Blood pressure
4	Diabetes	If Person having diabetes or not Yes=1,No=0
5	Alcohol	If Alcohol consume or not yes=1, no=0.
6	Genetic	If Genetically affected or not yes=1,no=0.
7	Stress	If Stress level if the person having stress or not yes=1 no=0.
8	Exercise	Person doing exercise or not if yes=1,no=0.
9	Valve block	Blocks in heart
10	Max heart rate	Maximum heart in 1min
11	Rest heart rate	Rest heart rate in 1min

Table 2: Attribute data set for heart disease

3.1. Pre Processing

Data pre-processing is a data mining approach that requires changing an unique data into a sensible association. Genuine data is frequently inadequate, variable and lacking in specific practices and it includes more blunders. Data pre-processing is a show system for performing such issues. Data pre-processing prepares essential data for moreover getting ready. There are numerous numbers of data pre-processing techniques. They are

- Data cleaning
- Data Integration
- Data reduction
- Data Transformation

The proposed system of Pre-processing utilized normalization. Normalization may build the accuracy and performance of mining calculations including split-up measures. The below figure 2 represent the raw data of the data set.

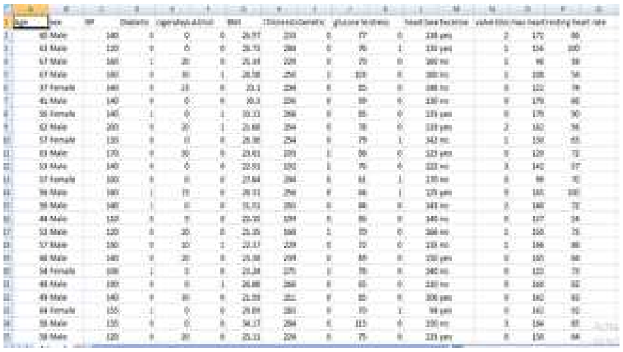
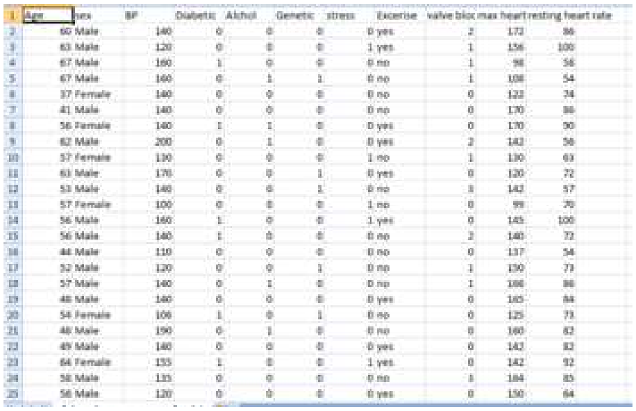


Fig 2: Raw Data

The figure 3 represents the pre processed data after the pre processed in using the different methods.



1	Age	sex	BP	Diabetic	Alchol	Genetic	stress	Excerise	valve bloc	max heart	resting heart rate
2	60	Male	140	0	0	0	0	yes	2	172	86
3	63	Male	120	0	0	0	1	yes	1	156	100
4	67	Male	160	1	0	0	0	no	1	98	56
5	67	Male	160	0	1	1	0	no	1	108	54
6	37	Female	140	0	0	0	0	no	0	122	74
7	41	Male	140	0	0	0	0	no	0	170	86
8	56	Female	140	1	1	0	0	yes	0	170	90
9	62	Male	200	0	1	0	0	yes	2	142	56
10	57	Female	130	0	0	0	1	no	1	130	63
11	63	Male	170	0	0	1	0	yes	0	120	72
12	53	Male	140	0	0	1	0	no	3	142	57
13	57	Female	100	0	0	0	1	no	0	99	70
14	56	Male	160	1	0	0	1	yes	0	143	100
15	56	Male	140	1	0	0	0	no	2	140	72
16	44	Male	110	0	0	0	0	no	0	137	54
17	52	Male	120	0	0	1	0	no	1	150	73
18	57	Male	140	0	1	0	0	no	1	166	86
19	48	Male	140	0	0	0	0	yes	0	165	84
20	54	Female	108	1	0	1	0	no	0	125	73
21	48	Male	190	0	1	0	0	no	0	160	62
22	49	Male	140	0	0	0	0	yes	0	142	82
23	64	Female	155	1	0	0	1	yes	0	142	92
24	58	Male	135	0	0	0	0	no	1	184	85
25	58	Male	120	0	0	0	0	yes	0	150	64

Fig 3: Pre-processed data

3.2. Methodology:

Today's, data mining plays an important part in the country. In this paper, data mining has been developed to improve the execution developing highlight choice approach.

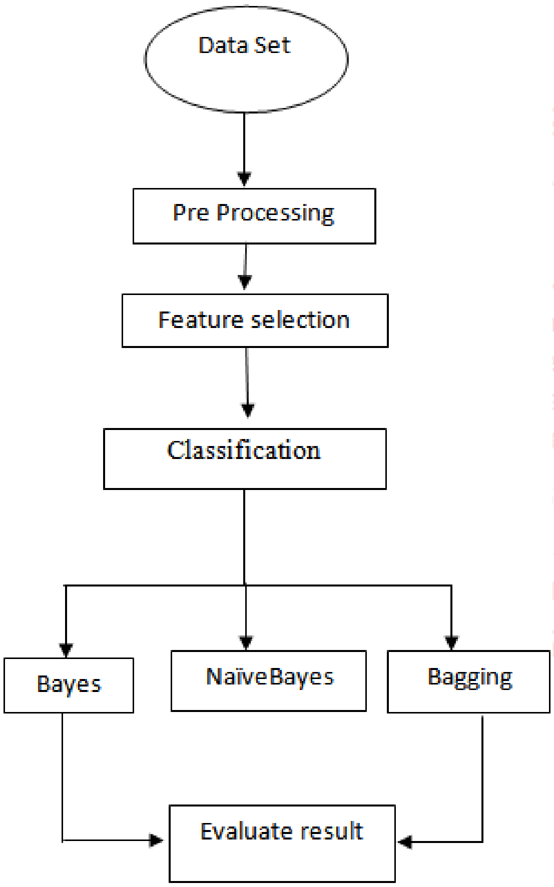


Fig 4: Frame work for NBASB

3.3. Feature selection:

Feature selection is similarly avoided as variable selection or attributes selection. For feature selection choice, various techniques are available, for example correlation analysis, random forest, relief and so on. In feature selection approach, choosing the appropriate properties and ordering of the insignificant qualities, different component determination approaches are connected to the pre-processed dataset.

3.3.1. Classification

Classification is a characteristic data mining technique used to categorize each item in a set of data into one of the predefined set of classes or groups. Classification is a technique makes use of mathematical techniques such as decision trees, Naïve Bayes, linear programming, neural network and statistics etc.

3.3.2. Naive Bayes

Naive Bayes technique is based on the 'Bayes Theorem'. Bayes theorem describes the probability of an event based on previous conditions which are relatin to particular event. e.g.. if Cardiovascular disease (CVD) are related to cholesterol level, then by using Bayes theorem, cholesterol level of person can be used more accurately for assessing the probability of CVD's. In machine learning, Naïve Bayes classifiers are treated as trouble free probabilistic classifiers and it is independent from assumptions between the features. [5]

3.3.3. Attribute Selection:

Attribute subset Selection is a technique which is used for data reduction in data mining process. Data reduction decreases the size of data so that it can be used for scrutiny purposes more proficiently. The dataset may have a huge number of attribute But some of those attributes can be unrelated or dismissed.

Methods of Attribute Subset Selection

1. Stepwise Forward Selection.
2. Stepwise Backward Elimination.
3. Combination of Forward Selection and Backward Elimination.

3.3.4. Bagging:

Bootstrap Aggregation famously knows as bagging, is a powerful and simple ensemble method.

An ensemble method is a technique that combines the predictions from many machine learning algorithms together to make more reliable and accurate predictions than any individual model. It means that we can say that prediction of bagging is very strong.

The main purpose of using the bagging technique is to improve Classification Accuracy.

- ➢ Accuracy Estimation
- ➢ Sampling with replacement
- ➢ Some may not be used, other may be used more than once.[6]

4. PERFORMANCE METRICS

The metrics used for the research work is labeled in this segment

There are four potential values of a confusion matrix for the two outcome values (Positive and Negative).

- ➢ True Positive (TP): TP is the total number of positive
- ➢ predictive instances which are classified correctly and
- ➢ truly positive.
- ➢ True Negative (TN): TP is the total number of negative
- ➢ predictive instances which are classified correctly and
- ➢ truly negative.
- ➢ False Positive (FP): FP is the total number of positive
- ➢ predictive instances which are classified incorrectly and
- ➢ truly negative.
- ➢ False Negative (FN): FN is the total number of negative predictive instances which are classified incorrectly and truly positive.[7]

4.1. Precision

Precision is the part of important illustrations between the retrieved instances. The Eq. of precision is given in Eq. (2)

$$Precision = TP/\ (TP+FP) \qquad (1)$$

4.2. Recall

Recall is the small part of suitable instances that have been reclaimed over the total amount of related instances. The Eq. of recall is given in Eq. (2).

$$\text{Recall} = TP/\ (TP + FN) \qquad (2)$$

4.3. F-Measure

The f-score (or f-measure) is considered based on the two times the precision times recall divided by the sum of precision and recall. The equation of F-Measure is given in Eq. (3).

$$\text{MCC}= \frac{\text{TP× TN -FP ×FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (3)$$

4.4. Roc Area

Roc Curves are commonly used to show in a graphical way the linking involving clinical understanding and specificity for every potential remove for a test or a procedure of tests.

4.5. Prc Area

The Precision-recall curves are not compressed by the count of patients without disease and with low test results. It is particularly suggested to use precision-recall curves as a enhancement to the regularly used ROC curves to attain the full picture when estimating and equating.

5. EXPERIMENTAL RESULT

5.1. WEKA

Weka is a tool that contains a collecting of representation implements and calculations for information examination and perceptive validating, together with graphical UIs for simple admittance to these functions. Weka boost a few standard data mining tasks, all the more particularly, Data pre-processing, classification, clustering, association, representation, besides highlight purpose. WEKA is a conventional data mining tool. It is utilized to interruption the most huge elements causing emptiness. It is likewise used to perform measurable examination of every individual distinguishing.

In experimental analysis we used classification techniques namely Naïve Bayes, Bagging and Attribute Selection to predict the classification accuracy of heart disease dataset. 3 experiment we implemented for the accuracy check first we executed Naïve Bayes with 11 attributes, next second experiment we executed Naïve Bayes with attribute selection and its executes 7 attributes and then third experiment NaïveBayes attribute removal, Bagging and applying test cross validation with percentage split.

6. PERFORMANCE ANALYSIS

Table3 represent the value of correlation analysis of different algorithms. The algorithms are TPRate, FPRate, Precision, Recall, F-Measure, MCC, ROC Area, PRC area.

Fig 5 shows the graphical representation of the correlation analysis of different algorithms that give more performance in Naives Bayes, Bagging, Attribute selection and percentage splits as 97%

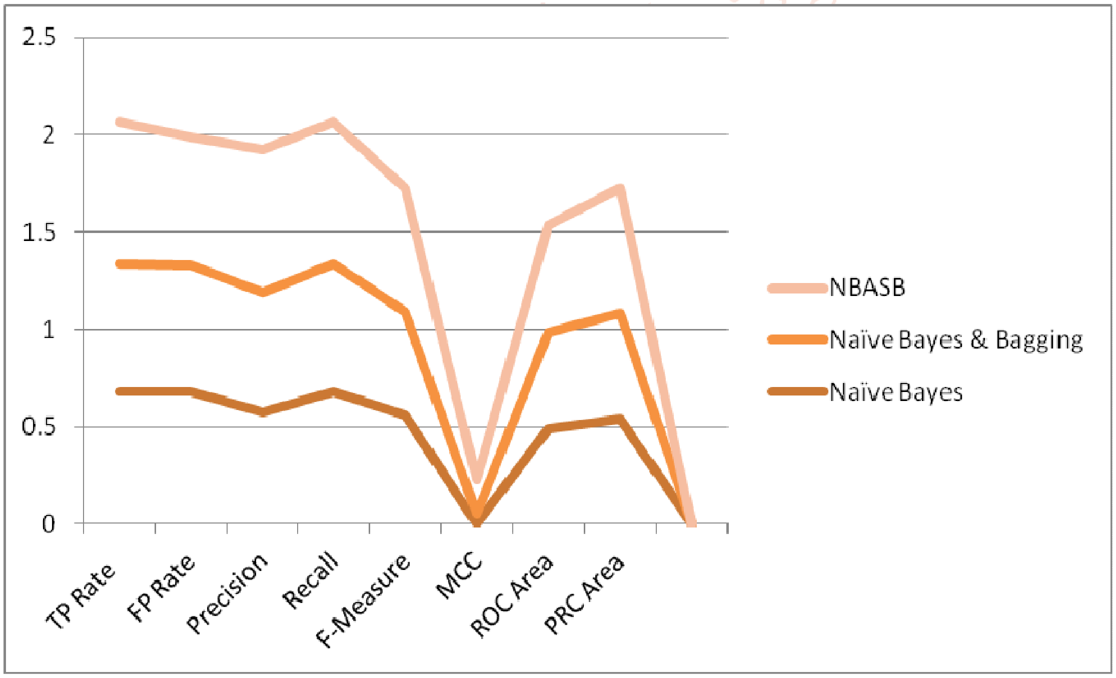


Fig 5: Graphical Representation of correlation analysis using algorithms

Table 3: Correlation Analysis of Algorithms

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Naïve Bayes	0.685	0.685	0.576	0.685	0.560	0.004	0.493	0.541
NaïveBayes & Bagging	0.655	0.644	0.614	0.655	0.531	0.047	0.492	0.543
NBASB	0.728	0.661	0.734	0.728	0.638	0.178	0.554	0.640

Table 4 Represent the time consuming time and accuracy check for different algorithms

Algorithms	Correctly Classified Instances	Incorrectly classified Instances	Time Taken	Mean Absolute Error	Relative Absolute Error
Naïve Bayes	68.48	31.51	0.25	43.25	99.18
Naïve Bayes &Bagging	65.54	34.46	0.24	45.11	99.75
NBASB	72.79	27.20	0.03	42.93	98.78

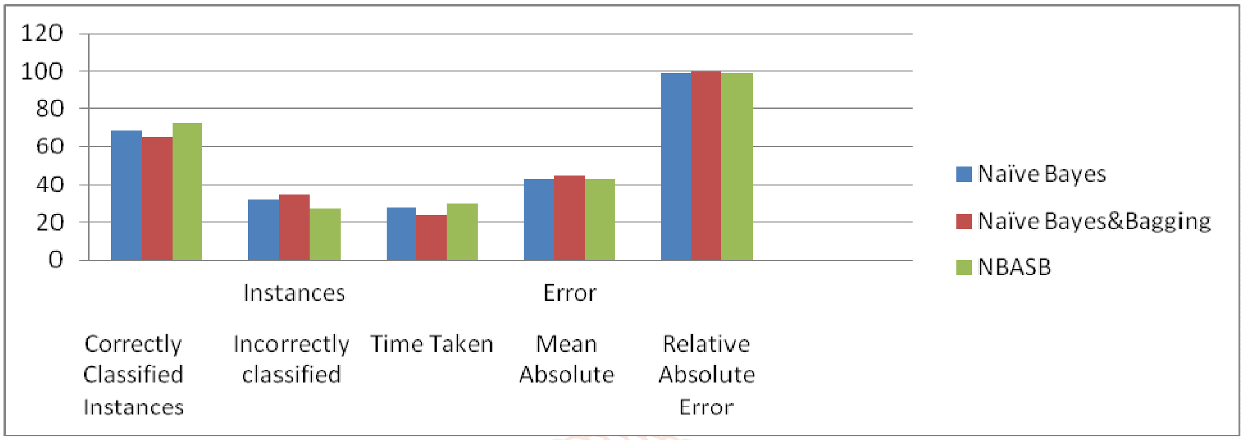


Figure 5: represent the graphical representation of accuracy and time taken.

7. CONCLUSION

In this paper the proposed a new algorithm for predicting the coronary diseases from medical records of Patient. We use Various Classification techniques they are Naïve Bayes, Bagging and Feature selection. We proposed algorithm Naïve Bayes, with other Meta algorithms like Bagging and attribute selection(NBASB) with the test mode has the percentage split 97% it increase the accuracy and time for predicting Coronary diseases.

REFERENCE

[1] Salha M. Alzahani, AfnanAlthopity, AshwagAlghamdi, BoushraAlshehri, and SuheerAljuaid "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction", Lecture Notes on Information Theory Vol. 2, No. 4, December 2014.

[2] H. Benjamin Fredrick David, S. Antony Belcy "Heart Disease Prediction Using Data Mining Techniques" ISSN: 2229-6956 (Online) ICTACT Journal On Soft Computing, October 2018, Volume: 09, Issue: 01 Doi: 10.21917/ijsc.2018.0253.

[3] Kanika Pahwa, Ravinder Kumar "Prediction of Heart Disease Using Hybrid Technique For Selecting Features"2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)GLA University, Mathura, Oct 26-28, 2017

[4] Shahed Anzarus Sabab, Ahmed Iqbal Pritom, Md. Ahadur Rahman Munshi, Shihabuzzaman," Cardiovascular Disease Prognosis Using Effective Classification and Feature Selection Technique".

[5] Mehdi Khundmir Iliyas, Vikas Maral "Heart Disease Prediction Using Naive Bayes And K-Means Techniques" Novateur Publications International Journal of Research Publications in Engineering and Technology [IJRPET] ISSN: 2454-7875 Volume 3, Issue 6, Jun.-2017

[6] https://t4tutorials.com/bagging-and-bootstrap-in-data-mining-machine-learning.

[7] Marjia Sultana, Afrin Haider, Mohammad ShorifUddin" Analysis of Data Mining Techniques for Heart Disease Prediction" 978-1-5090-2906-8/16/$31.00 ©2016 IEEE.