International Journal of
Intelligent Systems and
Applications in Engineering

ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

# The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods #

Enes Celik [1*], Muhammet Atalay [2], Adil Kondiloglu [3]

*Accepted 3rd September 2016*

***Abstract:*** Chronic kidney disease is a prolonged disease that damages the kidneys and prevents the normal duties of the kidneys. This disease is diagnosed with an increase of urinary albumin excretion lasting more than three months or with significant reduction in a kidney functions. Chronic kidney disease can lead to complications such as high blood pressure, anemia, bone disease and cardiovascular disease. In this study we have been investigated to determine the factors that decisive for early detection of chronic kidney disease, launching early patients treatment processes, prevent complications resulting from the disease and predict of disease. The study aimed diagnosis and prediction of disease using the data set that composed of data of 250 patients with chronic kidney disease and 150 healthy people. First, the chronic kidney disease data was classified with machine learning algorithms and then training and test results were analysed. The estimation results of chronic kidney disease were compared with similar data and studies.

***Keywords:*** *Chronic Kidney Disease, Machine Learning, Classification.*

## 1. Introduction

Today, technology is progressing day by day and the entry into almost every aspect of life and the data is obtained and stored in various areas. There are examples in every area of life which data is available and stored such as to keep track of income and expenses of the companies, to storage of information of students in school, to keep information that obtain as a result of personal work. Although often created large data sets used to store data, it is possible to convert data into information through a variety of operations on data and predict the future by the available data. Generally, techniques used for the conversion of the data according to the purpose information processing is known as data mining. By the application in daily life of information obtained data mining large gains can be achieved [1].

One of the areas where the application of data mining is the health field. In this area, determining in advance of possible illness, identifying the different cases where the disease associated with the forward-looking assumptions and patient conditions is possible by the studies on the generated data.

The job of the kidneys is to filter the blood. All the blood in your body passes through the kidneys several times a day. Kidneys has a responsibility to throw waste from the body, to control fluid balance and to regulate the electrolyte balance. The system is not working properly kidney stones and kidney failure may occur here. Factors which increase the risk of kidney disease are diabetes, hypertension, smoking, obesity, heart disease, kidney disease in the family, alcohol, drug abuse, drug overdose, age, race, sex, symptoms of kidney disease, urinary function changes, difficulty during urination, blood in the urine, back or back pain at the edges, fatigue, dizziness, lack of attention, always feeling cold, rash-itch, ammonia breath, metallic taste, nausea, vomiting and shortness of breath [2].

Chronic renal disease is a long-term disease that is being damaged kidneys and preventing them from doing actions which the normal duties of the kidneys such as cleaning the blood from harmful substances, the body maintain fluid balance, blood pressure regulation and hormone production. The disease is diagnosed with a significant reduction in urinary albumin excretion, increased or renal function for more than three months. This disease can lead to complications such as high blood pressure, anemia, bone disease and cardiovascular diseases [3].

Currently, there are many programs which hosts data mining algorithms. Some of these programs and software are Weka, R, Orange, Knime etc. These programs generally contain the same algorithms. But algorithms vary functioning forms of information because the algorithm authors are different and algorithms are developed. In our study, the support vector machine and decision tree algorithms of Weka program are used. Support vector machine is one of the simplest and effective methods used for classification. The decision tree algorithm is one of the machine learning methods and classification, is represented by tree branches and leaves forming a simple tree structure [4].

The study is to investigate the predictability of the disease. Using a variety data of chronic kidney patients and normal patients. In this study, we used a data set of 400 persons obtained from the "UCI Machine Learning Repository" data warehouse which is including the information of patients and people without [5].

[1] *Kirklareli University, Department of Computer Programming, Kirklareli, Turkey.*

[2] *Kirklareli University, Department of Quantitative Methods, Kirklareli, Turkey.*

[3] *Beykent University, Department of Computer Programming, Istanbul, Turkey.*

* *Corresponding Author: Email: enes.celik@klu.edu.tr*

# *Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.*

## 2. Literature Review

Vijayaranihas forecast kidney disease by using Navie Bayes and support vector machine algorithms. Mainly in the research, focused on finding the best classification algorithm according to the classification accuracy and execution time performance factors. It was found that the performance of the support vector machine algorithm to be better than Naive Bayes classifier from the experimental results [6].

Srinivasa has developed a dialysis support system with a kidney failure data and decision tree algorithm [7].

Kaladharhas achieved 97% success with J48 algorithm and 98% success with Random Forest algorithm in recognition of kidney stones with Weka [8].

Zadehhas achieved 80.85% success with WJ48 algorithm and 85.11% success with W Simple Cart algorithm for early detection of dialysis with Weka [9].

Hyarihas used the decision tree algorithm in the practice of chronic renal disease with Weka [10].

Song, has achieved 80% success with decision tree algorithm for renal failure disease using Weka [11].

Kumarhas achieved 96% success in identifying kidney stones using multi-layered network structure neural network algorithm with [12].

## 3. Methodology

### 3.1. Dataset

A dataset that used is composed of data of 24different variables that obtained from 250 patients with chronic kidney and obtained from 150 patients with non within 2 months from a hospital. The presence of 24 various tests measure to each patient is important in this data set in terms of demonstrating the applicability of conclusions derived from the results of studies to daily life. These data and properties are shown in Table 1.

**Table 1.** Patient Data and Units

| |
|---|
| 1-age-age |
| 2-bp-blood pressure |
| 3-sg-specific gravity |
| 4-al-albumin |
| 5-su-sugar |
| 6-rbc-red blood cells |
| 7-pc-pus cell |
| 8-pcc-pus cell clumps |
| 9-ba-bacteria |
| 10-bgr-blood glucose random |
| 11-bu-blood urea |
| 12-sc-serum creatinine |
| 13-sod-sodium |
| 14-pot-potassium |
| 15-hemo-hemoglobin |
| 16-pcv-packed cell volume |
| 17-wc-white blood cell count |
| 18-rc-red blood cell count |
| 19-htn-hypertension |
| 20-dm-diabetes mellitus |
| 21-cad-coronary artery disease |
| 22-appet-appetite |
| 23-pe-pedal edema |
| 24-ane-anemia |
| 25-class-class |

As shown in Table 1, patient data is also included age, blood pressure and appetite information in addition to various information of assays. The values and units of this data varies according to the type of data.1, 2, 10, 11, 12, 13, 14, 15, 16, 17 and 18 of the variables in data are get units specified numerical values; 3, 4 and 5 of the variables in data are get specific numerical values; and others consist of the specified non-numeric values and the initial values are written in parenthesis next to the data. There is also a 25thvariable in the data of the person that to hold the information “ckd” if a person patient and “notckd” if a person not the patient.

### 3.2. Support Vector Machine

There are many algorithms can be used for classification in Weka software. Support vector machine is one of the very effective and simple methods that used for classification. Simply, support vector machine is an algorithm that working the method of to draw a line in the plane between the two groups and the separation of these two groups. In Figure 1, separating the groups from each other by the algorithm is shown in Figure 1.

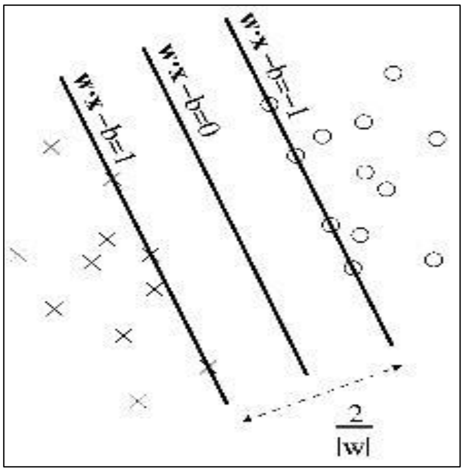


**Figure 1.** Separation of the Group by the Support Vector Machine [13]

In Figure 1 the first, verify are drawn two groups identified the closest element to each other, representing these elements, for to separate the two groups placed in a two-dimensional plane by support vector machine. The process of separating two different groups from each other is maintained for determining the correct point on the equidistant two lines drawn between the two towards. In the Weka, SMO (Sequential Minimal Optimization) algorithm is using like the support vector machine algorithm [14].

### 3.3. Decision Tree

In the decision tree algorithm method, the class labels are level of the tree leaves, leading to the leaves and the process on start arms are expressed forming a tree structure [15]. The tree structure as a result of the decision tree algorithm is shown in Figure 2.

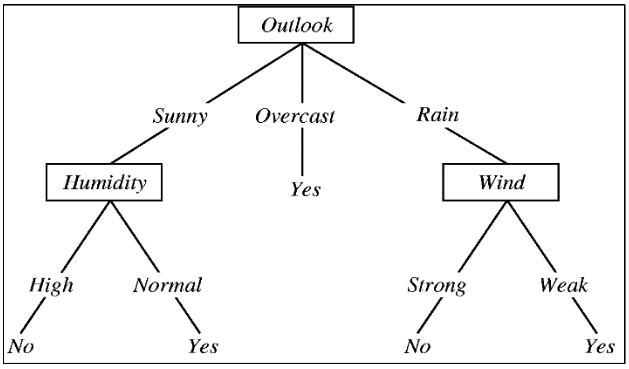


**Figure 2.** Decision Tree Algorithm [16]

As the decision tree algorithm, J48 algorithm of Weka program was used based on basic ID3 and C4.5 algorithms. The entropy is used for determine which qualifications based on branching when creating a decision tree. Branch are determined by calculated the gain metrics for each attribute.  Because of the branching start from the highest gain criterion qualifications, it is understood that which the classification criterion is more effective after formation of tree.

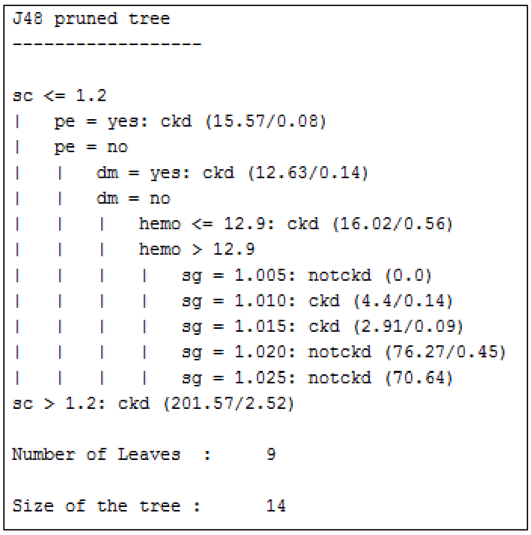
## 4. Test Results

J48 and SMO algorithms used in the estimation process is performed by trained testing and training data in two different sizes. As Test-1, algorithms are trained by creating data in approximately 66% of all data. The estimation procedure is done for algorithms by creating test data with the remaining 34% of data. As Test-2, algorithms are trained by creating data in approximately 10% of all data. The estimation procedure is done for algorithms by creating test data with the remaining 90% of data. The results of the prediction procedure are shown in Table 2.

**Table 2.** J48 and SMO Algorithms Test Results

| Result Name | J48 | |
|---|---|---|
| | Test-1 | Test-2 |
| Percentage of Correct Classification | 100 % | 91.6667 % |
| Kappa Statistic | 1 | 0.827 |
| Mean Absolute Error | 0.0226 | 0.1547 |
| Mean Square Error | 0.0837 | 0.2753 |
| Relative Absolute Error | 4.8492 % | 34.292  % |
| Absolute Relative Root Square Error | 17.4751 % | 55.3729 % |
| **Result Name** | **SMO** | |
| | **Test-1** | **Test-2** |
| Percentage of Correct Classification | 97.0588 % | 96.1111 % |
| Kappa Statistic | 0.9368 | 0.9184 |
| Mean Absolute Error | 0.0294 | 0.0389 |
| Mean Square Error | 0.1715 | 0.1972 |
| Relative Absolute Error | 6.3003 % | 8.6192 % |
| Absolute Relative Root Square Error | 35.7951 % | 39.6706 % |

Analysing the test results in Table 2; for both algorithms; it was found that the error values of from Test-1 results are lower than the error values of from Test-2. Because of the data size used to train the algorithm in Test-1 is larger than the data size of the Test-2 process; algorithms are better educated, forecast accuracy higher than Test-2 and forecast transactions are also lower as a result of the error values in the Test-1 processing. Therefore it the magnitude of the training data affects estimation process. Kappa statistic values are also outside the error values in the table. To be close to 1of this value indicates of the accuracy of the estimation process. The reliability and validity of the estimate decreases when the kappa statistic closer to zero. The decision tree formed in the Test-1 results of J48 algorithm is shown in Figure 3.



```
J48 pruned tree
------------------


sc <= 1.2
|   pe = yes: ckd (15.57/0.08)
|   pe = no
|   |   dm = yes: ckd (12.63/0.14)
|   |   dm = no
|   |   |   hemo <= 12.9: ckd (16.02/0.56)
|   |   |   hemo > 12.9
|   |   |   |   sg = 1.005: notckd (0.0)
|   |   |   |   sg = 1.010: ckd (4.4/0.14)
|   |   |   |   sg = 1.015: ckd (2.91/0.09)
|   |   |   |   sg = 1.020: notckd (76.27/0.45)
|   |   |   |   sg = 1.025: notckd (70.64)
sc > 1.2: ckd (201.57/2.52)

Number of Leaves  :     9

Size of the tree :      14
```

**Figure 3.** Decision Tree Formed in the Test-1 Results of J48 Algorithm

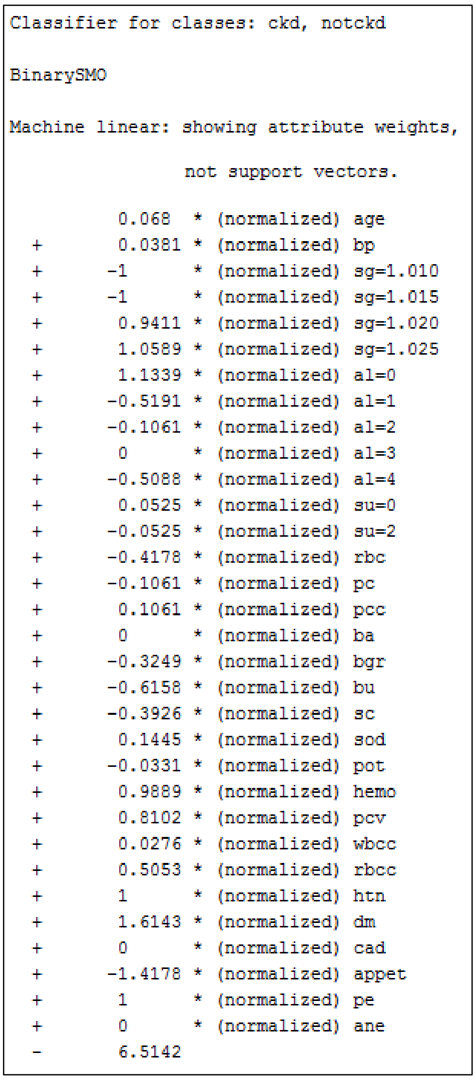
When examined the decision tree in Figure 3; it is observed that begins with the capability of serum creatinine (sc). Because the leaves of decision trees are ranked according to earnings criteria, it can say that the first decisive feature in the estimation process is this feature. Other features that make up the branches of the decision tree are also of importance for the estimation process, respectively. Each leaf is divided into branches by >, <,> = and <= operations. The last leaf of the branches, it is reaching the (ckd) value or (unckd) value if the people are patients or not patients. The first number listed in parentheses next to this value illustrates the total weight of examples in the leaves, the second number indicates the weight of the incorrect classification. In addition, in the estimates, the total number of leaves of the created tree (Number of Leaves), and the size of the tree (Size of the Tree) are also indicated. Also a decision tree created in the Test-2 results and a decision tree is formed in the Test-1 result are the same. This is also shows that the size of the training data don't changing characteristics that impact on estimates but the size of the data affects education level of learning algorithm.

The results of estimation procedures of SMO algorithm, the weight vector which will be made of the best estimate process is determined by creating a weighting coefficient for each feature. In Figure 4, SMO algorithm weights as a result of the features that generate Test-1 operations are shown.

In Figure 4, when the weight of the vector by the result of SMO algorithm examined, it has been observed that some features of the weight is calculated as zero. As it is understood from this result, zero weight value properties does not have an effect on the process of the estimating. The characteristics which its coefficient equal zero has no effect on the estimation process. Because all the features form the weight vector with coefficient values, the features having nonzero coefficients are not possible to sort by the coefficient values.

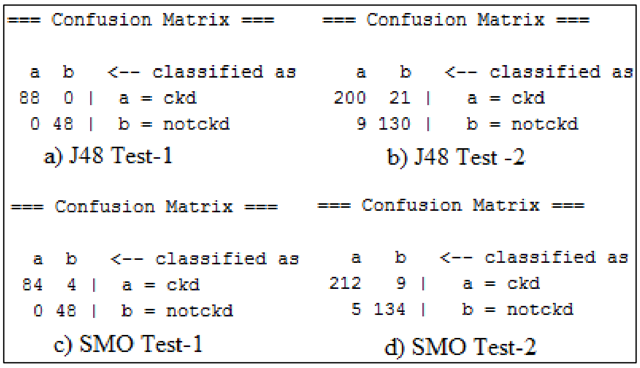
Weight vectors generated as a Test 1 and Test-2 result of SMO algorithm are the same. This is also showed that effective to change properties on the estimated size of the data affect the level of training algorithm. It has been show that the data size cannot change the properties that effective on the estimates but it showed that the algorithm affect the level of education.

Confusion Matrix of J48 and SMO algorithms formed in Test-1 and Test-2 results are shown in Figure 5.



```
Classifier for classes: ckd, notckd

BinarySMO

Machine linear: showing attribute weights,

                 not support vectors.

           0.068  * (normalized) age
  +        0.0381 * (normalized) bp
  +       -1      * (normalized) sg=1.010
  +       -1      * (normalized) sg=1.015
  +        0.9411 * (normalized) sg=1.020
  +        1.0589 * (normalized) sg=1.025
  +        1.1339 * (normalized) al=0
  +       -0.5191 * (normalized) al=1
  +       -0.1061 * (normalized) al=2
  +        0      * (normalized) al=3
  +       -0.5088 * (normalized) al=4
  +        0.0525 * (normalized) su=0
  +       -0.0525 * (normalized) su=2
  +       -0.4178 * (normalized) rbc
  +       -0.1061 * (normalized) pc
  +        0.1061 * (normalized) pcc
  +        0      * (normalized) ba
  +       -0.3249 * (normalized) bgr
  +       -0.6158 * (normalized) bu
  +       -0.3926 * (normalized) sc
  +        0.1445 * (normalized) sod
  +       -0.0331 * (normalized) pot
  +        0.9889 * (normalized) hemo
  +        0.8102 * (normalized) pcv
  +        0.0276 * (normalized) wbcc
  +        0.5053 * (normalized) rbcc
  +        1      * (normalized) htn
  +        1.6143 * (normalized) dm
  +        0      * (normalized) cad
  +       -1.4178 * (normalized) appet
  +        1      * (normalized) pe
  +        0      * (normalized) ane
  -        6.5142
```

**Figure 4.** SMO Algorithm Weight Vector Form of Test-1 Results



```
=== Confusion Matrix ===        === Confusion Matrix ===

  a  b   <-- classified as        a   b   <-- classified as
 88  0 |  a = ckd               200  21 |   a = ckd
  0 48 |  b = notckd              9 130 |   b = notckd
```

a) J48 Test-1 b) J48 Test -2

```
=== Confusion Matrix ===        === Confusion Matrix ===

  a  b   <-- classified as        a   b   <-- classified as
 84  4 |  a = ckd               212   9 |   a = ckd
  0 48 |  b = notckd              5 134 |   b = notckd
```

c) SMO Test-1 d) SMO Test-2

**Figure 5.** Confusion Matrixes of J48 and SMO algorithms

As it is seen from the confusion matrix of algorithms; J48 classification algorithm was not making a mistake in Test-1and the error is 0%. However, 21 patients were classified as non-patients and 9 non-patients were also classified as patient and the error is 8.33% in Test-2. SMO algorithm were classified 4 patients as non-patients in the Test-1 results and error is 2.94%. And 9 patients were classified as non-patients and 5 non-patients were also classified as patient and the error is 3.88% in Test-2.

## 5. Conclusions

Chronic Kidney Disease a longer disease that prevents the normal duties of the kidneys and causing any damage the kidneys. The early detection of this disease is very important in terms of health and treatment costs. In this study, using the data of consisting the 250 chronic kidney disease patients and 150 non-patient people data set, classification of patients was estimated by the help of support vector machine and decision tree algorithm. Training and testing process of algorithms were measured by creating two different sets of data. As a result of different data size and prediction operations performed with different algorithms, it has been shown that the size of the training data algorithms to be largely effective for the estimating.

In the classification stage, decision tree has been more successful than the support vector machine recognition of 97% to 100% recognition. Forecasting and phase accuracy, it has been observed that made an accurate estimate of 100% rate with the decision tree in the Test-1 data set and an accurate estimate of 91.67% rate in the Test-2 data set. Also it has been observed that made an accurate estimate of 97.06% rate with the SVM in the Test-1 data set and an accurate estimate of 96.12% rate in the Test-2 data set. With the data used in this study showed that the decision tree gives better results than the support vector machine for the early diagnosis of chronic kidney disease. High rate results was obtained in terms of performance as compared with results in the literature.

## References

[1] A.S. Albayrak and Ş.K. Yılmaz, "Veri Madenciliği Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama", Süleyman Demirel Üniversitesi İİBF Dergisi, Volume 14, 2009.

[2] S. Bala and K. Kumar, "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique", International Journal of Computer Science and Mobile Computing, Volume 3, Issue 7, 2014.

[3] G. Süleymanlar, Akdeniz Üniversitesi Tıp Fakültesi İç Hastalıkları Nefroloji Bilim Dalı, Online Accessed: May, 2016, www.medikalakademi.com.tr/kronik-bobrek-yetmezligi-baslangic-belirtileri-tani-tedavisi/, 2013.

[4] G. Silahtaroğlu, "Veri Madenciliği Kavram ve Algoritmaları", Papatya Yayıncılık Eğitim, İstanbul, 2013.

[5] Soundarapandian, Online Accessed: May, 2016, https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease, 2015.

[6] S. Vijayarani and S. Dhayanand, "Data mining classification algorithms for kidney disease prediction", International Journal on Cybernetics & Informatics (IJCI), Vol. 4, No. 4, 2015.

[7] S. R. Raghavan, V. Ladik and K. B. Meyer, "Developing decision support for dialysis treatment of chronic kidney failure", IEEE Transactıons on Information Technology in Biomedicine, vol. 9, no. 2, 2005.

[8] K. Krishna, A. Rayavarapu and V. Vadlapudi, "Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis", Open Access Scientific Reports, Volume 1, Issue 12, 2012.

[9] M. K. Zadeh, M. Rezapour and M. M. Sepehri, "Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients", International journal of hospital research, Volume 2, Issue 1, 2013.

[10] Y. Abeer and A. Hyari, "Chronic Kidney Disease Prediction System Using Classifying Data Mining Techniques", Library of University of Jordan, 2012.

[11] X. Song, Z. Qiu and J. Mu, “Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field”, International Journal of Advancements in Computing Technology (IJACT), Volume 4, Number 3, 2012.

[12] K. Kumar, "Artificial neural networks for diagnosis of kidney stones disease." International Journal of Information Technology and Computer Science (IJITCS) Volume 4, Issue 7, 2012.

[13] Ş.E. Şeker, “İş Zekası ve Veri Madenciliği”, Cinius Yayınları, İstabul, Türkiye, 2013.

[14] E. Celik and A. Kondiloglu, "Detection of fake banknotes with Artificial Neural Networks and Support Vector Machines", 23th Signal Processing and Communications Applications Conference (SIU), 2015.

[15] Y. Özkan, “Veri Madenciliği Yöntemleri”, Papatya Yayıncılık Eğitim, Türkiye, 2013.

[16] J.R. Quinlan, Online Accessed: May, 2016, http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm, 2015.