**Project proposal: Medical prediction for patients of kidney disease**

1. **Motivation and problem definition**

Almost a large number of the population worldwide is affected with a major health problem, chronic kidney disease. As a result, early detection and characterization are considered to be critical factors in the management and control of this long-lasting kidney disease. These tasks have been traditionally performed by well-trained healthcare professionals; however, they are still some of the most challenging work due to the subtle signs and difficult to detect symptoms hidden in data set. Herein, use of well-organized data mining techniques is shown to expose hidden information from clinical and laboratory patient data, which can be helpful to assist physicians in maximizing accuracy for identification of disease severity stage. The existing works without the use of the machine learning algorithms fail to provide the accuracy of prediction to the needed extent. So, our project will try to indicate that applying different machine learning algorithms provide better classification and prediction performance for determining whether one patient has chronic kidney disease. The project will try to predict the chronic kidney diseases of patients using systematic and automatic methodologies. Among the methodologies, the machine learning algorithm and feature selection are some of the very kinds.

**II.         Proposed solution**

To give a solution for the challenge of medical prediction for kidney disease patients, we will work though  3 consecutive phases:

1. **Choosing and preprocessing dataset**

We choose to use the  [Chronic\_Kidney\_Disease Dataset](https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease?fbclid=IwAR2bJXrFFo9VK)  because it’s openly available online for public access on the Machine Learning Repository website[\*]. The data was retrieved from an Indian hospital over approximately a 2-month period. It contains 25 attributes and 400 records. The first 24 attributes are the risk factors of chronic kidney disease while the 25th attribute is the classification of the disease (ckd or notckd). There are 11 numeric attributes and 14 nominal attributes. The numeric attributes include Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count. The nominal attributes include Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell clumps, Bacteria, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia, Class.

﻿The order of the attributes, the attributes’ name, the attributes’ datatype and attributes unit is given as follows:

      1.Age(numerical)

                  age in years

      2.Blood Pressure(numerical)

                   bp in mm/Hg

      3.Specific Gravity(nominal)

                  sg - (1.005,1.010,1.015,1.020,1.025)

      4.Albumin(nominal)

                  al - (0,1,2,3,4,5)

      5.Sugar(nominal)

                  su - (0,1,2,3,4,5)

      6.Red Blood Cells(nominal)

                  rbc - (normal,abnormal)

      7.Pus Cell (nominal)

                  pc - (normal,abnormal)

      8.Pus Cell clumps(nominal)

                  pcc - (present,notpresent)

      9.Bacteria(nominal)

                  ba  - (present,notpresent)

      10.Blood Glucose Random(numerical)

                  bgr in mgs/dl

      11.Blood Urea(numerical)

                  bu in mgs/dl

      12.Serum Creatinine(numerical)

                  sc in mgs/dl

      13.Sodium(numerical)

                  sod in mEq/L

      14.Potassium(numerical)

                  pot in mEq/L

      15.Hemoglobin(numerical)

                  hemo in gms

      16.Packed  Cell Volume(numerical)

      17.White Blood Cell Count(numerical)

                  wc in cells/cumm

      18.Red Blood Cell Count(numerical)

                  rc in millions/cmm

      19.Hypertension(nominal)

                  htn - (yes,no)

      20.Diabetes Mellitus(nominal)

                  dm - (yes,no)

      21.Coronary Artery Disease(nominal)

                  cad - (yes,no)

      22.Appetite(nominal)

                  appet - (good,poor)

      23.Pedal Edema(nominal)

                  pe - (yes,no)

      24.Anemia(nominal)

                  ane - (yes,no)

      25.Class (nominal)

                  class - (ckd,not ckd)

This dataset has rich resourceful information  and each record is the statistic about the medical situation of an individual. However, this dataset contains a considerable amount of missing values and mistyped characters, a preprocessing data phase needs to be implemented to overcome this drawback.

1. **Features extraction**
   1. **Motivation**

This is an efficient data preprocessing technique in data mining for reducing dimensionality of data. In medical diagnosis, it is very important to identify most significant risk factors related to disease. Relevant feature identification helps in the removal of unnecessary, redundant attributes from the disease dataset which, in turn, gives quick and better results. So, in our project we used feature selection, also known as Variable Selection, as an extensively used data preprocessing technique in data mining which we basically used for reduction of data by eliminating insignificant and superfluous attributes from our CKD dataset. Moreover, this technique enhances the comprehensibility of the data, facilitates better visualization of the data, reduces training time of learning algorithms and improves the performance of prediction.

In some related works, such as the research presented in (12) has considered 5 attributes: blood pressure, serum creatinine, packed cell volume, hyper- tension, and anemia to calculate the L-factor and clustered Chronic Kidney Disease (CKD) and non-CKD patients based on the L-factor value. According to their evaluation CKD cannot be detected based on their L-factor classifiers. Other works (5), (16) have evaluated machine learning algorithms such as back propagation neural networks, radial basis functions, random forests and SVMs and achieved up to 85.3% accuracy on identifying CKD. Also, (22) performs feature selection techniques such as information gain, gain ratio, or attribute evaluation and fusion based feature selection to identify relevant features, but their evaluation has not presented the relevant selected features. Hence, in our project, we tried to use another approach.

Because the original dataset has 24 attributes and not all of them give relevant information about the medical situation of an individual. If we keep the irrelevant attributes, they may cause some noise to the predicted results. Also, because the considered attributes come from the patients, which reflect their health situation, there must be some correlations between some attributes. Keeping all the attributes will lead to the existence of redundant information. Another important reason to have the feature extraction/selection come from the interpretability property of Machine Learning models. Medical experts need to have a clear view about the strong predictors as the medical metrics so that they can analyze and explain the situation to the patients. As such, this phase is important.

* 1. **Methods**

There exist numerous applications of relevant feature identification techniques in the healthcare sector. Filter methods, wrapper methods, ensemble methods and embedded methods are some of the popularly used techniques.

* **Univariate Feature Extraction**: This technique involves more of a manual kind of work. Visiting every feature and checking its importance with the target. There are some great tricks that you should keep under your sleeve to implement Univariate Feature Selection. Having a proper domain knowledge and checking the variance are some good examples we used in our data feature selection.
* **PCA**: The main purposes of a PCA are the analysis of the data to identify patterns and finding patterns to reduce the dimensions of the dataset with minimal loss of information. It will try to reduce dimensionality by exploring how one feature of the data is expressed in terms of the other features (linear dependency). Feature selection instead, takes the target into consideration. PCA works best on a dataset having 3 or higher dimensions. Because, with higher dimensions, it becomes increasingly difficult to make interpretations from the resultant cloud of data.
* **Wrapper method (Recursive Feature Elimination- RFE):** It works by recursively removing attributes and building a model on those attributes that remain. It uses an external estimator that assigns weights to features (for example, the coefficients of a linear model) to identify which attributes (and the combination of attributes) contribute the most to predicting the target attribute.
* **Embedded method (select K-best**): The SelectKBest class just scores the features using a function and then "removes all but the k highest scoring features
* **Filter method (E.g. Chi-Square):** It is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.
  1. **Result**

**﻿**We used SelectKBest, PCA, and Recursive Feature Elimination (RFE) to determine the best attributes -strong predictors. We argue that, if an attribute is a strong predictor, it should appear in the result of various Feature Extraction techniques, therefore by doing experiments with various aforementioned algorithms, we will detect which attributes appear frequently and then select them.

The following set is the set of 10 best attributes that we obtained after running 3 algorithms:

* 𝑠𝑒𝑡1 = [𝑤𝑏𝑐𝑐, 𝑏𝑢, 𝑏𝑔𝑟, 𝑎𝑙, 𝑠𝑐, 𝑝𝑐𝑣, 𝑠𝑢, ℎ𝑡𝑛, 𝑝𝑐, 𝑑𝑚]

**A screenshot of a cell phone

Description automatically generated**

**﻿**The ﻿ExtraTreesClassifier is then used to determine the feature importance set.

**A screenshot of a cell phone

Description automatically generated**

* **﻿**𝑠𝑒𝑡2 = [bp, rbc, 𝑏𝑔𝑟, dm, sg, sc, pcv, pc, al, htn]

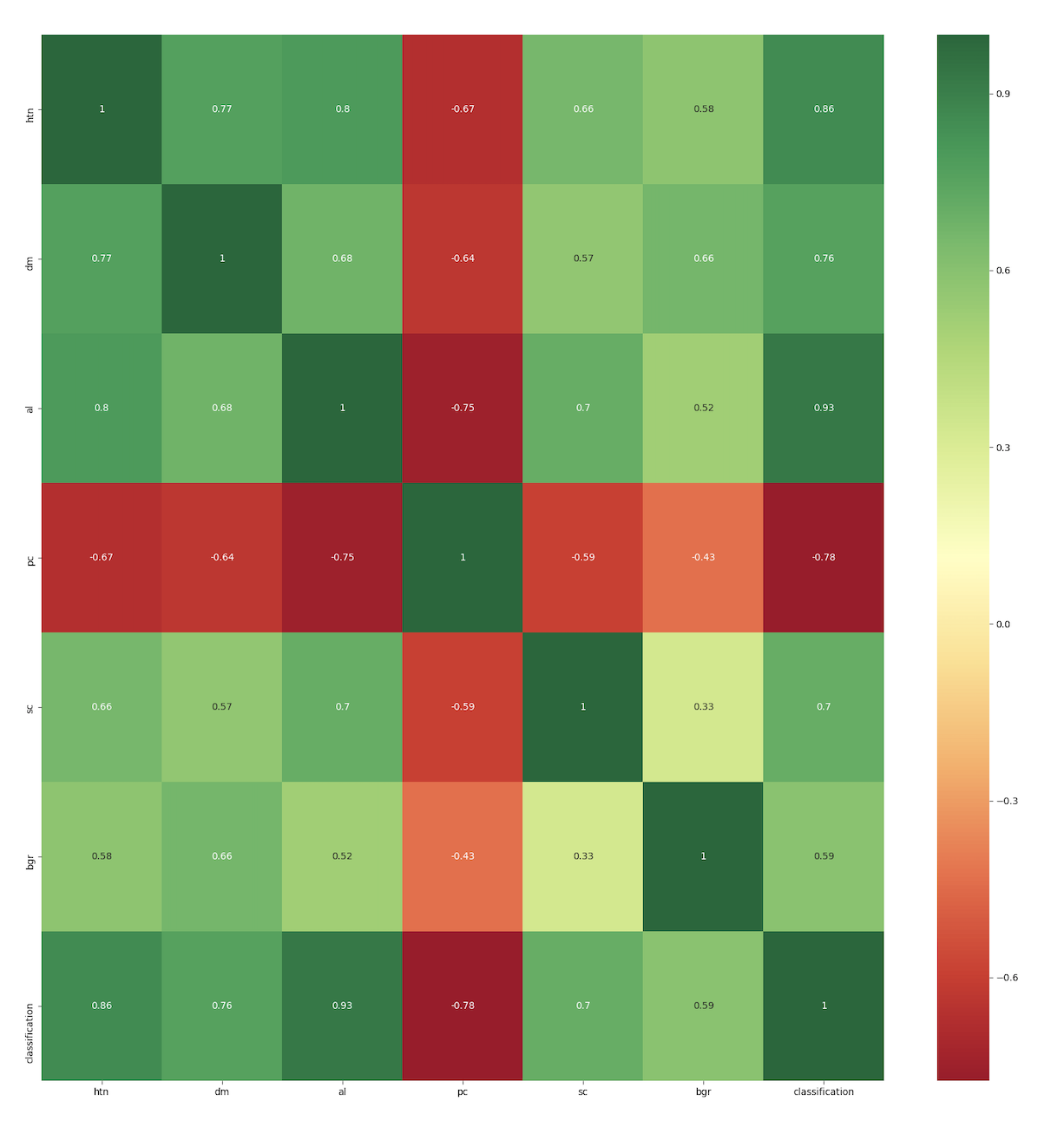
﻿𝑆𝑡𝑟𝑜𝑛𝑔 𝐴𝑡𝑟𝑟𝑖𝑏𝑢𝑡𝑒𝑠 = 𝑠𝑒𝑡1 ∩ 𝑠𝑒𝑡 2 = [′𝑝𝑐𝑣′, ′ℎ𝑡𝑛′, ′𝑑𝑚′, ′𝑎𝑙′, ′𝑝𝑐′, ′𝑠c′, ′𝑏𝑔𝑟′]

We have the heat map of original features:

A picture containing colorful, white

Description automatically generated

﻿Correlations between these chosen attributes are obtained below:



3.     **Apply various machine learning models with trial-and-test strategy**

After the preprocessing and features extraction phases,  we will apply various machine learning models to determine which model will give the best result in predicting kidney disease patients. The model we will do the experiments are: ﻿Decision Tree Classifier, ﻿k-Nearest Neighbor Classifier, ﻿Logistic regression Classifier, ﻿Support Vector Machine Classifier, ﻿Nonlinear Support Vector Machine, ﻿Ensemble Methods

After getting the results, we will compare the efficiency of the models to determine which one is the best choice. Also, we probably propose some improvement for our approach with Neural Network.

﻿[\*] Machine Learning Repository - Center for Machine Learning and Intelligent Systems.

Retrived from

<https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease?fbclid=IwAR2bJXrFFo9VK>

**References**

1. Z. Masetic and A. Subasi, “Congestive heart failure detection using random forest classifier,” *Computer Methods and Programs in Biomedicine,* vol. 130, pp. 54-64, 2016.

[Google Scholar](https://scholar.google.com/scholar?q=Z.%20Masetic%20and%20A.%20Subasi%2C%20%E2%80%9CCongestive%20heart%20failure%20detection%20using%20random%20forest%20classifier%2C%E2%80%9D%20Computer%20Methods%20and%20Programs%20in%20Biomedicine%2C%20vol.%20130%2C%20pp.%2054-64%2C%202016.)

1. E. Alickovic, J. Kevric and A. Subasi, “Performance evaluation of EMD, DWT, and WPD for automated epileptic seizure detection and prediction,” *Submitted to Biomedical Signal Processing and Control,* 2016.

[Google Scholar](https://scholar.google.com/scholar?q=E.%20Alickovic%2C%20J.%20Kevric%20and%20A.%20Subasi%2C%20%E2%80%9CPerformance%20evaluation%20of%20EMD%2C%20DWT%2C%20and%20WPD%20for%20automated%20epileptic%20seizure%20detection%20and%20prediction%2C%E2%80%9D%20Submitted%20to%20Biomedical%20Signal%20Processing%20and%20Control%2C%202016.)

1. M. J. Pérez-Sáeza, . D. Prieto-Alhambra, C. Barrios, M. Crespo, D. Redondo, X. Nogués, A. Díez-Pérez and J. Pascual, “Increased hip fracture and mortality in chronic kidney disease individuals: The importance of competing risks,” *Bone,* vol. 73, p. 154–159, 2015.

[Google Scholar](https://scholar.google.com/scholar?q=M.%20J.%20P%C3%A9rez-S%C3%A1eza%2C%20.%20D.%20Prieto-Alhambra%2C%20C.%20Barrios%2C%20M.%20Crespo%2C%20D.%20Redondo%2C%20X.%20Nogu%C3%A9s%2C%20A.%20D%C3%ADez-P%C3%A9rez%20and%20J.%20Pascual%2C%20%E2%80%9CIncreased%20hip%20fracture%20and%20mortality%20in%20chronic%20kidney%20disease%20individuals%3A%20The%20importance%20of%20competing%20risks%2C%E2%80%9D%20Bone%2C%20vol.%2073%2C%20p.%20154%E2%80%93159%2C%202015.)

1. A. M. Cueto-Manzano, L. Cortes-Sanabria, H. R. Martinez-Ramirez, E. Rojas-Campos, B. Gomez-Navarro and M. Castillero-Manzano, “Prevalence of Chronic Kidney Disease in an Adult Population,” *Archives of Medical Research,* vol. 45, pp. 507-513, 2014.

[Google Scholar](https://scholar.google.com/scholar?q=A.%20M.%20Cueto-Manzano%2C%20L.%20Cortes-Sanabria%2C%20H.%20R.%20Martinez-Ramirez%2C%20E.%20Rojas-Campos%2C%20B.%20Gomez-Navarro%20and%20M.%20Castillero-Manzano%2C%20%E2%80%9CPrevalence%20of%20Chronic%20Kidney%20Disease%20in%20an%20Adult%20Population%2C%E2%80%9D%20Archives%20of%20Medical%20Research%2C%20vol.%2045%2C%20pp.%20507-513%2C%202014.)

1. P. Sinha and P. Sinha, “Comparative study of chronic kidney disease prediction using knn and svm,” *International Journal of Engineering Research and Technology*, vol. 4, no. 12, 2015.
2. A. Levin and P. E. Stevens, “Summary of KDIGO 2012 CKD Guideline: behind the scenes, need for guidance, and a framework for moving forward,” *Kidney International,* vol. 85, no. 1, p. 49–61, 2014.

[Google Scholar](https://scholar.google.com/scholar?q=A.%20Levin%20and%20P.%20E.%20Stevens%2C%20%E2%80%9CSummary%20of%20KDIGO%202012%20CKD%20Guideline%3A%20behind%20the%20scenes%2C%20need%20for%20guidance%2C%20and%20a%20framework%20for%20moving%20forward%2C%E2%80%9D%20Kidney%20International%2C%20vol.%2085%2C%20no.%201%2C%20p.%2049%E2%80%9361%2C%202014.)

1. Z. Chen, Z. Zhang, R. Zhu, Y. Xiang and P. B. Harrington, “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” *Chemometrics and Intelligent Laboratory Systems,* vol. 153, p. 140–145, 2016.

[Google Scholar](https://scholar.google.com/scholar?q=Z.%20Chen%2C%20Z.%20Zhang%2C%20R.%20Zhu%2C%20Y.%20Xiang%20and%20P.%20B.%20Harrington%2C%20%E2%80%9CDiagnosis%20of%20patients%20with%20chronic%20kidney%20disease%20by%20using%20two%20fuzzy%20classifiers%2C%E2%80%9D%20Chemometrics%20and%20Intelligent%20Laboratory%20Systems%2C%20vol.%20153%2C%20p.%20140%E2%80%93145%2C%202016.)

1. P. Muthukumara and G. S. S. Krishnan, “A similarity measure of intuitionistic fuzzy soft sets and its application in medical diagnosis,” *Applied Soft Computing,* vol. 41, p. 148–156, 2016.

[Google Scholar](https://scholar.google.com/scholar?q=P.%20Muthukumara%20and%20G.%20S.%20S.%20Krishnan%2C%20%E2%80%9CA%20similarity%20measure%20of%20intuitionistic%20fuzzy%20soft%20sets%20and%20its%20application%20in%20medical%20diagnosis%2C%E2%80%9D%20Applied%20Soft%20Computing%2C%20vol.%2041%2C%20p.%20148%E2%80%93156%2C%202016.)

1. “UCI Machine Learning Repository: Chronic Kidney Disease (CKD) Data Set,” 15 November 2016. [Online]. Available: view-source:<https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease>#. [Accessed Nov 15 2016].
2. B. Widrow, D. E. Rumelhart and M. A. Lehr, “Neural networks: applications in industry, business and science,” *Communications of the ACM,* vol. 12, pp. 93-105, 1994.

[Google Scholar](https://scholar.google.com/scholar?q=B.%20Widrow%2C%20D.%20E.%20Rumelhart%20and%20M.%20A.%20Lehr%2C%20%E2%80%9CNeural%20networks%3A%20applications%20in%20industry%2C%20business%20and%20science%2C%E2%80%9D%20Communications%20of%20the%20ACM%2C%20vol.%2012%2C%20pp.%2093-105%2C%201994.)

1. S. Armin, Data Mining and Knowledge Discovery Handbook, 2nd ed., O. Maimon and L. Rokach, Eds., New York: Springer, 2010.

[Google Scholar](https://scholar.google.com/scholar?q=S.%20Armin%2C%20Data%20Mining%20and%20Knowledge%20Discovery%20Handbook%2C%202nd%20ed.%2C%20O.%20Maimon%20and%20L.%20Rokach%2C%20Eds.%2C%20New%20York%3A%20Springer%2C%202010.)

1. A. Dubey, “A classification of ckd cases using multivariate k-means clustering.” *International Journal of Scientific and Research Publica- tions (IJSRP)*, vol. 5, August 2015.
2. L. Rokach and O. Maimon, Data Mining and Knowledge Discovery Handbook, 2nd ed., M. Oded and R. Lior, Eds., New York: Springer, 2010.

[Google Scholar](https://scholar.google.com/scholar?q=L.%20Rokach%20and%20O.%20Maimon%2C%20Data%20Mining%20and%20Knowledge%20Discovery%20Handbook%2C%202nd%20ed.%2C%20M.%20Oded%20and%20R.%20Lior%2C%20Eds.%2C%20New%20York%3A%20Springer%2C%202010.)

1. L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, Wadsworth Int. Group, 1984.

[Google Scholar](https://scholar.google.com/scholar?q=L.%20Breiman%2C%20J.%20Friedman%2C%20R.%20Olshen%20and%20C.%20Stone%2C%20Classification%20and%20Regression%20Trees%2C%20Wadsworth%20Int.%20Group%2C%201984.)

1. J. R. Quinlan, “Induction of Decision Trees,” *Machine Learning,* vol. 1, pp. 81-106, 1986.

[Google Scholar](https://scholar.google.com/scholar?q=J.%20R.%20Quinlan%2C%20%E2%80%9CInduction%20of%20Decision%20Trees%2C%E2%80%9D%20Machine%20Learning%2C%20vol.%201%2C%20pp.%2081-106%2C%201986.)

1. S.RamyaandN.Radha,“Diagnosisofchronickidneydiseaseusingma- chine learning algorithms,” *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 1, 2016.
2. J. R. Quinlan, C4.5: Program for Machine Learning, CA, Morgan Kaufman Publishing, 1993.

[Google Scholar](https://scholar.google.com/scholar?q=J.%20R.%20Quinlan%2C%20C4.5%3A%20Program%20for%20Machine%20Learning%2C%20CA%2C%20Morgan%20Kaufman%20Publishing%2C%201993.)

1. L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, p. 5–32, 2001.

[Google Scholar](https://scholar.google.com/scholar?q=L.%20Breiman%2C%20%E2%80%9CRandom%20Forests%2C%E2%80%9D%20Machine%20Learning%2C%20vol.%2045%2C%20p.%205%E2%80%9332%2C%202001.)

1. E. Alickovic and A. Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest,” *Neural Computing and Applications,* pp. 1-11, 2015.

[Google Scholar](https://scholar.google.com/scholar?q=E.%20Alickovic%20and%20A.%20Subasi%2C%20%E2%80%9CBreast%20cancer%20diagnosis%20using%20GA%20feature%20selection%20and%20Rotation%20Forest%2C%E2%80%9D%20Neural%20Computing%20and%20Applications%2C%20pp.%201-11%2C%202015.)