“Forecasting Chronic Kidney Disease using the Decision tree and Random Forest Algorithms in Machine learning”

Miss. Rekha A. Shidnekoppa
Assistant Professor
Bindu Nagesh Hotagar, Sindhu H Kumachagi, Ganesh P Badi, and Katyayani N Dandin
Department of Computer Science and Engineering
Tontadarya College of Engineering Gadag, Karnataka, India

Abstract–One of the most significant global health problems is chronic kidney disease. It has a high morbidity and death rate. Patients frequently miss the diagnosis of CKD because there are no symptoms in the early stage of the condition. The early diagnosis of CKD in the critical state enables patients to get immediate treatment, which slows the disease’s further development. The application of the machine learning techniques in health care for disease categorization and prediction has increased as a result of the availability of pathology data. The classification of CKD using a machine learning model by SVM (Support Vector Machine) proposed in this paper. We trained around 24 features of datasets among 19 features were tested and achieved 93.75% accuracy.

Index Terms–chronic kidney disease (CKD), Support Vector Machine (SVM), Machine Learning (ML).

I. INTRODUCTION (HEADING 1)

Kidney sickness has formed into a typical disease with difficult issues. A different gathering of sicknesses influencing the design and usefulness of the urinary organ is alluded to as kidney infection. It is all perceived that a little variety in the structure and capability of the urinary organs could increase the probability of issues in other organ frameworks. Approximately 10% of the world’s population suffers from chronic kidney disease (CKD), which cause a large number of passing each year. The primary causes of the disease are listed as being Many illnesses, including obesity and hypertension, can result in CKD

Family history: if you have a history of renal illness, dialysis, or kidney transplantation, you may be at an increased risk of developing the condition.

Medicines: Certain medications, such as over-the -counter pain relievers, can either induce or exacerbate renal disease.

Age and race: renal disease may be more common in elderly adults and in some racial groupings. The early kidney disease diagnosis protects the patient from life-threatening complications. The factors that cause renal illness must be properly examined in order to forecast them.

II. LITERATURE SURVEY

Using data mining algorithms in the Hadoop, Guneet Kaur suggested a system for forecasting the CKD in 2017[1]. They employ two KNN and SVM-based data mining classifier. Here the manually chosen data columns are used to do the prediction analysis. In this system, SVM classifier provides better accuracy than KNN.

Implementing machine learning technique [2] such as the Multi-Layer Perceptron Algorithm, Logistic Regression, Decision Trees, Nave Bayes, Support Vector Machine, and K-Nearest Neighbor Algorithm, Baisakhi Chakraborty proposed developing a CKD prediction system in 2019. These are used and the effectiveness of each is evaluated in relation to the outcomes for accuracy, precision, and recall. The system was finally implemented using Random Forest.

Siddheshwar Tekale presented an AI framework in 2018 that makes use of Choice tree SVM techniques [3]. After comparing the results of the two approaches, it is concluded that SVM produces the best results. It has a less laborious forecast cycle the enables specialists to dissect patients more quickly.

A system that made use of the Random Forest algorithm and the Back propagation neural network was proposed by J. Snegha in 2020[4]. In this case, they compared the two algorithms and discovered that the Back Propagation algorithm produces the best results because it employs the feed forward neural network, a supervised learning network.

III. PROPOSED SYSTEM

Constant kidney illness (CKD) is one of the primary explanations for death all through the world nowadays. The expression "ongoing kidney sickness" means persevering through harm to the kidneys that can crumble long term. If the harm is entirely awful, then, at that point, kidney might stop working. This is called End stage renal disappointment. The expectation of CKD is maybe the most critical and testing issues in the clinical benefits assessment. To obtain concealed information from the given dataset, information mining is used to make the choices. This paper means to aid the forecast of constant kidney sickness (CKD) by using the help vector machine (SVM) classifier in the clinical space.

In this paper, we have investigated ML methods and done an exploratory examination into group phases of CKD. In this proposed framework, we have fabricated an ML model utilizing SVM to characterize whether a patient has CKD or not. Prior to applying arrangement calculation, we wiped out a couple of the highlights utilizing the highlight choice technique.

(1) Advantages of the proposed system

It is noted that existing studies have obtained the lowest accuracy; while the proposed system has obtained an accuracy of 94% with the proposed SVM. Finally, it is observed that the proposed has optimal results compared with existing systems. The proposed system results show that SVM is considered as the best classifier when contrasted with other classifier algorithms.

(2) Algorithms Used

- *Support Vector Machine (SVM) model:* SVM or Sponsorship Vector Machine is a straight model for gathering and backslide issues. It can deal with both straight and indirect problems. SVM is a straightforward possibility. The calculation generates a hyperplane or line that divides the data into classes.
- *Random forest:* It is an order calculation comprising of numerous choices of trees. Irregular woodland calculation utilizes arbitrary subsets of elements and thus, it can perform very well with a high layered dataset.
- *Decision tree:* A choice tree is a regulated learning calculation that is not parametric and can be used for both order and relapse tasks. It has a tree structure with different levels, with a root hub, branches, interior hubs, and leaf hubs.

(3) Modules:

- *Data collection*: This is the primary genuine step towards the genuine improvement of an AI model, gathering information. This is a basic step that will flow to how great the model will be, the more and better information that we get; the better our model will perform. There are a few procedures to gather the information, similar to web scratching, manual mediation and so forth.

- *Data set*: The dataset consists of 401 individual data. There are 25 columns in the dataset, which are described below.
  Age, bp, sg, al, su, rbc, pc, pcc, ba, bgr, bu, se, sod, pot, hemo, pcv, wc, rc, htn, dm, cad, appet, pe, ane, class.

- *Data preparation:* Gather data and get it ready for training. Eliminate duplicates, correct errors, manage missing qualitie, standardize, and transform information into new types, among other things.) Randomize information, which eliminates the impact of the specific request for which we collected or, possibly, prearranged our data. Visualize information to help make connections between factors or class lopsided characteristics that are applicable (inclination alert!), or carry out additional exploratory investigation into the preparation and evaluation sets.

- *Model selection*: We carried out the help vector machine calculation after obtaining an exactness of 0.9375 on the test set. Support Vector Machines (SVM) are learning frameworks that utilization a speculation space of direct capabilities in a high layer including space, prepared with a gaining calculation from a streamlining hypothesis that carries out a gaining predisposition got from a measurable learning hypothesis. The objective of the SVM is to find the ideal hyperplane that partitions the two classes. There can be various planes that can separate the two classes, yet the primary spotlight is on figuring out such a plane that we can accomplish the greatest edge between the classes. It implies picking the hyperplane with the goal of separation from the hyperplane to the closest information point is expanded. How can it function?

  We got familiar with the most common way of isolating the two classes with a hyper-plane.

  - *Scenario 1:* A, B, and C are the three hyperplanes we have here. Recognize the appropriate hyperplane right now to arrange the triangle and rectangle. To identify the appropriate hyperplane, you should keep the following rule in mind: Choose the hyperplane that better separates the two classes. Hyper-plane "B" has performed this work in an astonishing manner in these circumstance.

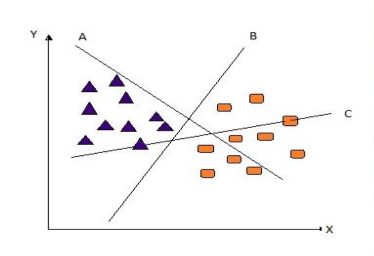


Fig 1: Identify the right hyper-plane (Scenario 1)

  - *Scenario 2:* To identify the correct hyper-plane, follow the previously discussed guidelines. The trick is that SVM selects the hyperplane that precisely groups the classes before expanding the edge, so some of you may have chosen hyperplane B because it has a higher edge than hyperplane A. Hyperplane B's order is off, but hyperplane A has all been grouped correctly. The correct hyperplane is thus, A.

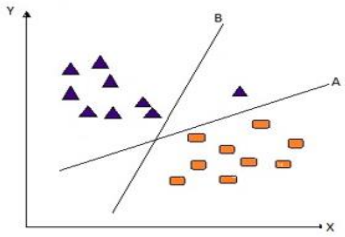


Fig 2: Identify the right hyper-plane (Scenario 2)

- *Scenario 3:* Underneath, I cannot confine the two classes using a straight line, as one of the triangles lies in the space of other(circle) class as a special case. One star at the opposite end resembles an exception for the triangle class, as I have proactively mentioned. A part of the SVM calculation looks for the hyperplane with the greatest edge and ignores exceptions. So, we can say that SVM grouping is strong against anomalies.

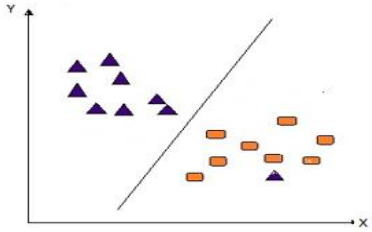


Fig 3: Identify the right hyper-plane (Scenario 3)

- *Analyse and prediction:* Here only 18 features are considered:
  Age, bp, al, su, rbc, pc, pcc, ba, bgr, bu, sc, pot, wc, htn, dm, cad, pe, ane, class.

- *Exactness on the test set:* We got an exactness of 0.9375% on test set.

- *Saving the Prepared Model:* When you are sufficiently sure to take your prepared and tried model into the ideal prepared climate, the initial step is to save it to a tie or h5. pkl document like a pickle that makes use of a library. Ensure you have pickle introduced in your current circumstance. Then, we should import the module and dump the model into the. pkl document.

- *Flowchart:*

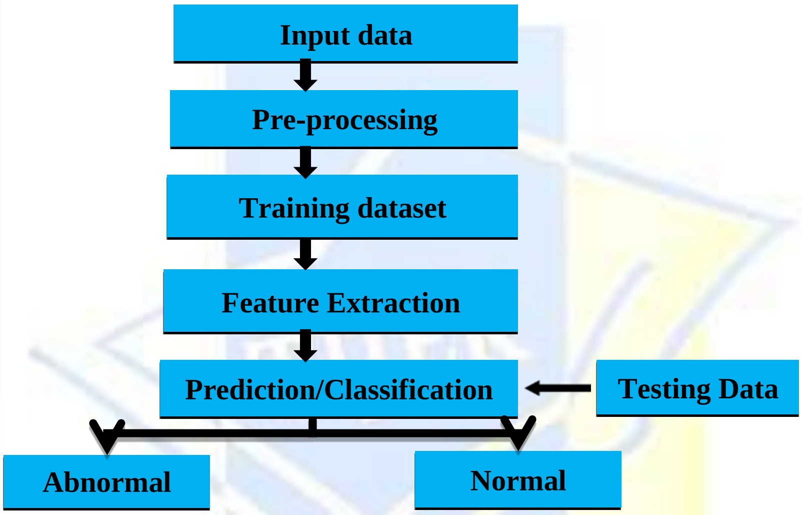


Fig 4: Process of forecasting chronic kidney disease

Input data: To give contributions to an AI model, you need to include the upsides of the elements you used to prepare your AI model. Then we can use it in a model.

Pre-processing: Information pre-processing in AI alludes to the strategy of getting ready (cleaning and coordinating) the crude information to make it reasonable for a structure and preparing AI models.

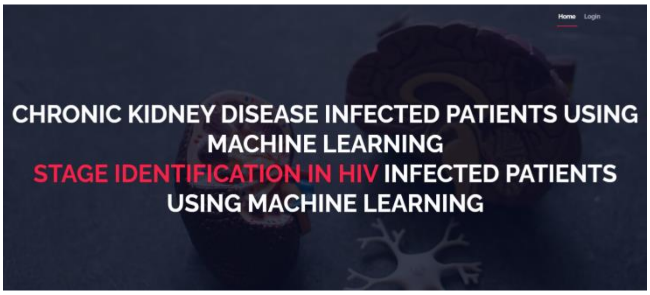
Training dataset: The preparation information is the greatest (in - size) subset of the first dataset, which is utilized to prepare or fit the AI model. The ML calculations take care of the preparation information first and foremost, enabling them to determine how to set expectations for the assigned assignment.

Feature extraction: The most common method of transforming crude information into mathematical elements that can be handled while protecting the data in the first informational index is referred to as highlight extraction. It produces better outcomes than applying AI solely to crude data.

Testing data: Testing information is utilized to actually look at the exactness of the model. The preparation dataset is, for the most part, bigger in size comparison with the testing dataset.

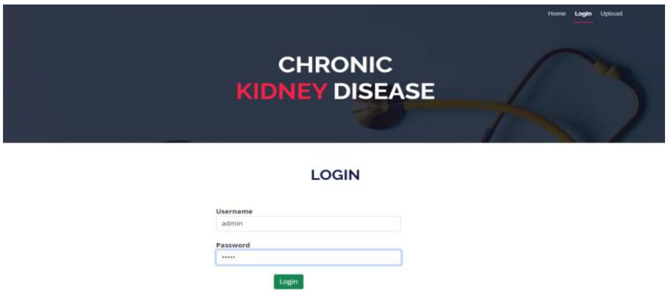
Classification: In order, the model is completely prepared utilizing the preparation information, and afterward it is assessed on test information prior to being utilized to perform a forecast on new concealed information.

IV. RESULTS



CHRONIC KIDNEY DISEASE INFECTED PATIENTS USING MACHINE LEARNING
STAGE IDENTIFICATION IN HIV INFECTED PATIENTS USING MACHINE LEARNING

Fig 5: Main page

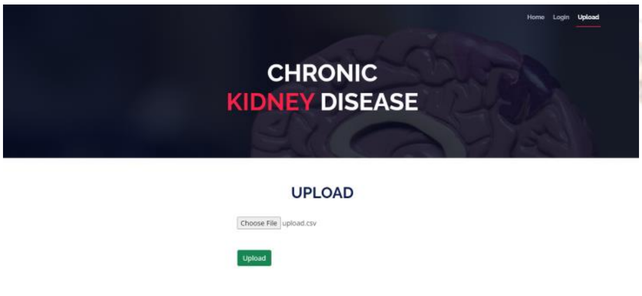


CHRONIC KIDNEY DISEASE

LOGIN

Username: admin
Password: •••••

Fig 6: Login Page

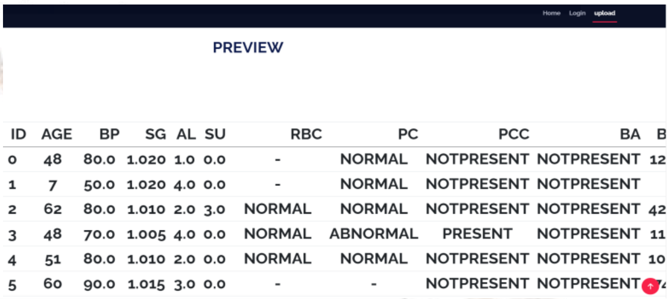


CHRONIC KIDNEY DISEASE

UPLOAD

Choose File upload.csv

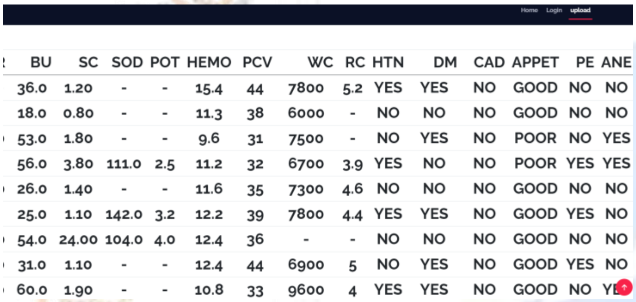
Fig 7: Uploading csv file



PREVIEW

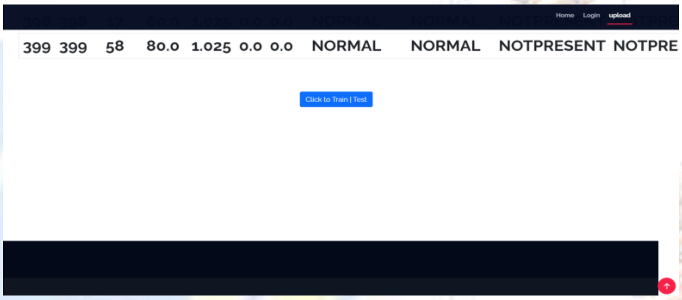
ID	AGE	BP	SG	AL	SU	RBC	PC	PCC	BA	B
0	48	80.0	1.020	1.0	0.0	-	NORMAL	NOTPRESENT	NOTPRESENT	12
1	7	50.0	1.020	4.0	0.0	-	NORMAL	NOTPRESENT	NOTPRESENT	
2	62	80.0	1.010	2.0	3.0	NORMAL	NORMAL	NOTPRESENT	NOTPRESENT	42
3	48	70.0	1.005	4.0	0.0	NORMAL	ABNORMAL	PRESENT	NOTPRESENT	11
4	51	80.0	1.010	2.0	0.0	NORMAL	NORMAL	NOTPRESENT	NOTPRESENT	10
5	60	90.0	1.015	3.0	0.0	-	-	NOTPRESENT	NOTPRESENT	7

Fig 8: Feature set 1



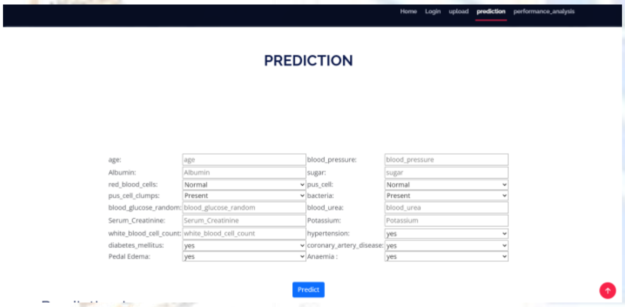
BU	SC	SOD	POT	HEMO	PCV	WC	RC	HTN	DM	CAD	APPET	PE	ANE
36.0	1.20	-	-	15.4	44	7800	5.2	YES	YES	NO	GOOD	NO	NO
18.0	0.80	-	-	11.3	38	6000	-	NO	NO	NO	GOOD	NO	NO
53.0	1.80	-	-	9.6	31	7500	-	NO	YES	NO	POOR	NO	YES
56.0	3.80	111.0	2.5	11.2	32	6700	3.9	YES	NO	NO	POOR	YES	YES
26.0	1.40	-	-	11.6	35	7300	4.6	NO	NO	NO	GOOD	NO	NO
25.0	1.10	142.0	3.2	12.2	39	7800	4.4	YES	YES	NO	GOOD	YES	NO
54.0	24.00	104.0	4.0	12.4	36	-	-	NO	NO	NO	GOOD	NO	NO
31.0	1.10	-	-	12.4	44	6900	5	NO	YES	NO	GOOD	YES	NO
60.0	1.90	-	-	10.8	33	9600	4	YES	YES	NO	GOOD	NO	YE

Fig 9: Feature set 2



399	399	58	80.0	1.025	0.0	0.0	NORMAL	NORMAL	NOTPRESENT	NOTPRE

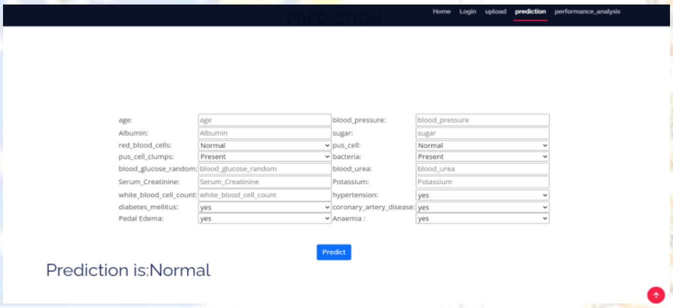
Fig 10: Training features



PREDICTION

age:	age	blood_pressure:	blood_pressure
Albumin:	Albumin	sugar:	sugar
red_blood_cells:	Normal	pus_cell:	Normal
pus_cell_clumps:	Present	bacteria:	Present
blood_glucose_random:	blood_glucose_random	blood_urea:	blood_urea
Serum_Creatinine:	Serum_Creatinine	Potassium:	Potassium
white_blood_cell_count:	white_blood_cell_count	hypertension:	yes
diabetes_mellitus:	yes	coronary_artery_disease:	yes
Pedal Edema:	yes	Anaemia :	yes

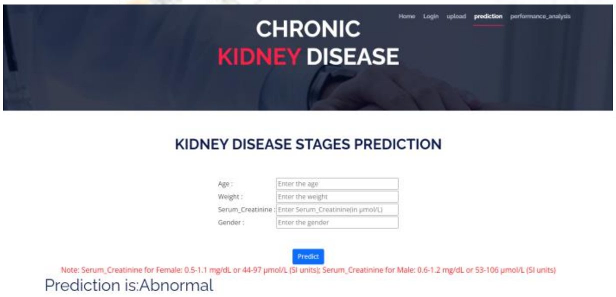
Fig 11: Inputting feature values



age:	age	blood_pressure:	blood_pressure
Albumin:	Albumin	sugar:	sugar
red_blood_cells:	Normal	pus_cell:	Normal
pus_cell_clumps:	Present	bacteria:	Present
blood_glucose_random:	blood_glucose_random	blood_urea:	blood_urea
Serum_Creatinine:	Serum_Creatinine	Potassium:	Potassium
white_blood_cell_count:	white_blood_cell_count	hypertension:	yes
diabetes_mellitus:	yes	coronary_artery_disease:	yes
Pedal Edema:	yes	Anaemia :	yes

Prediction is:Normal

Fig 12: Prediction is normal



CHRONIC KIDNEY DISEASE

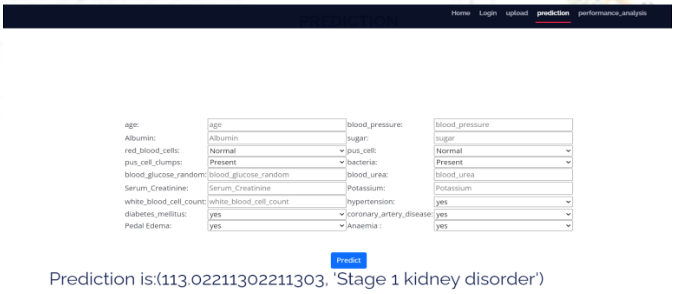
KIDNEY DISEASE STAGES PREDICTION

Age :	Enter the age
Weight :	Enter the weight
Serum_Creatinine :	Enter Serum_Creatinine(in μmol/L)
Gender :	Enter the gender

Note: Serum_Creatinine for Female: 0.5-1.1 mg/dL or 44-97 μmol/L (SI units); Serum_Creatinine for Male: 0.6-1.2 mg/dL or 53-106 μmol/L (SI units)

Prediction is:Abnormal

Fig 13: Prediction is abnormal



age:	age	blood_pressure:	blood_pressure
Albumin:	Albumin	sugar:	sugar
red_blood_cells:	Normal	pus_cell:	Normal
pus_cell_clumps:	Present	bacteria:	Present
blood_glucose_random:	blood_glucose_random	blood_urea:	blood_urea
Serum_Creatinine:	Serum_Creatinine	Potassium:	Potassium
white_blood_cell_count:	white_blood_cell_count	hypertension:	yes
diabetes_mellitus:	yes	coronary_artery_disease:	yes
Pedal Edema:	yes	Anaemia :	yes

Prediction is:(113.02211302211303, 'Stage 1 kidney disorder')

Fig 14: Stage of ckd

V. CONCLUSION

Classification of the stage of chronic kidney disease in HIV-infected patients is very helpful for both the patient and the doctor to make timely and accurate clinical decisions. We have used SVM to classify chronic kidney disease (CKD) in HIV patients in this paper. SVM has outperformed in the CKD classification, according to our study. In the future, medical image analysis and features-based SVM could be used together to support diagnosis based on various imaging modalities.

VI. REFERENCES

[1] Guneet Kaur, “Predict Chronic Kidney Disease using Data Mining in Hadoop,” 2017 International Conference on Inventive Computing and Informatics.

[2] Baisakhi Chakraborty,” Development of Chronic Kidney Disease Prediction Using Machine Learning,” 2019 International Conference on Intelligent Data Communication Technologies

[3] Siddheshwar Tekale, “Prediction of Chronic Kidney Disease Using Machine Learning,” 2018 International Journal of Advanced Research in Computer and Communication Engineering.

[4] J. Snegha,” Chronic Kidney Disease Prediction using Data Mining,” 2020 International Conference on Emerging Trend.

[5] N. V. G. Raju, K. P. Lakshmi, K. G. Praharshitha, and C. Likhitha, ‘‘Prediction of chronic kidney disease (CKD) using data science,’’ in Proc. Int. Conf. Intel. Compute. Control Syst. (ICCS), May 2019, pp. 642–647.

[6] A. Ogunleye and Q.-G. Wang, ‘‘XGBoost model for chronic kidney disease diagnosis,’’ IEEE/ACM Trans. Compute. Biol. Bioinf., vol. 17, no. 6, pp. 2131–2140, Nov. 2020.

TECHNIX INTERNATIONAL JOURNAL FOR ENGINEERING RESEARCH
TIJER
OPEN ACCESS JOURNAL