International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Chronic Kidney Disease (CKD) Phenotype and Its Association with Single Nuclotide Polymorpisms (SNPs) using Elastic Net Method

Maureen Zerlina Oktaviani[1], Angga Aditya Permana*[2], Analekta Tiara Perdana[3]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Chronic kidney disease (CKD) is one of critical disorders that typically has no symptoms until late stages. Early intervention in its initial stages can delay this disease progression. Single nucleotide polymorphism (SNP) can be used as a genetic variation that can be observed with CKD phenotype. Association study of SNP and CKD phenotype needs assistance from a machine learning algorithm that can process data with many dimensions. In this research, Elastic Net used as a method to associate SNP with phenotype for CKD. Based on the result, it showed that the Elastic Net method can select 88 significant SNP with a low MAE score that is 0.755 with a good explained variance (R2) which is 0.999. Also, there are significant SNP that were chosen with Elastic Net method such as Rhd, F2, Col1a1, Nos3, F5, Gypc, F8, F10, Bcam, Rbp4, Plg, Thpo, Myh9, Akt1, Fgb, and F7.

Keywords: *chronic kidney disease (CKD), elastic net, single nucleotide polymorphism (SNP).*

1. Introduction

One of the organs that is quite vital but often neglected by many people is the kidneys. Kidneys have an essential function in the human body as a blood filter from metabolic products so if the health of the kidneys is disturbed and cannot perform this function properly, the human will likely experience Chronic Kidney Disease (CKD).

CKD is health problem that globally increased and become the leading cause of death [1], especially in high-income countries as WHO report in 2016 [2]. In Indonesia, it is estimated that almost 70,000 patients have CKD [3]. Kidney disease treatment in Indonesia ranks second in the largest financing from BPJS Kesehatan under heart disease [4]. Thus, more up-to-date treatment techniques are required compared to those currently available to help treat CKD patients more effectively [5].

Early detection of CKD is essential to increase the effectiveness of treatments [1]. Precision medicine is a method of modern treatment that aims to develop prevention and treatment strategies based on environment, lifestyle, genetics, and individual phenotype diversity [6]. One of the strategies in question is to analyse the relationship between genotype (genetic code) and phenotype (observable characteristic) from an individual with a particular disease [7].

Precision medicine can be used by medical personnel to identify the stage of a patient's CKD by assessing kidney function. However, the application of direct research to humans is hindered by Ethics, Legality, and Social Implications (ELSI). Humans and mammals have systems of nearly the same complexity. one example of an animal is the rat which is the most common research model [8]. Scientists genetically modified mice by manipulating mouse genomes and gene function so that the rat kidneys can be used for kidney biology studies. Rat kidneys are morphologically similar to human kidneys [9]. Moreover, CKD should be detected with the least possible tests and at low cost. So, the objective of this research is to provide an effective model to predict the CKD by least number of predictors.

The most common form of genetic variation found in the human body is Single Nucleotide Polymorphism (SNP) [10]. SNP is a difference in the arrangement of single nucleotide bases in the genome of an individual which causes genetic variation in a population [11]. In addition, SNPs can also be used as biomarkers. Some studies have shown that SNPs in the ACE2 and TMPRSS2 regulatory sequences can influence mRNA expression and stability. SNPs also play a role together with other genes in finding the metabolism of a drug [12]. Various methods are used to look for an association between SNP genomes and certain disease phenotypes because these associations can yield information on which SNP significantly influences the phenotype that is being observed [13].

To identify CKD association analysis, it is necessary to collect SNP data based on the genetic variation of the patient. Association analysis using SNP data has its own challenge due to the large dimensions and numbers of small samples, so it is prone to curse of dimensionality problems [14]. To process data with many dimensions, it is necessary to reduce the number of dimensions without losing a lot of important information. one method is by feature selection or feature extraction or both [15].

[1,2] Department of Informatic, Faculty of Engineering and Informatic, Universitas Multimedia Nusantara, Indonesia
[3] Department of Biology, Faculty of Science, UIN Sultan Maulana Hasanuddin Banten, Indonesia
** Corresponding Author Email: angga.permana@umn.ac.id*

Previously, several similar studies have been carried out regarding SNP with phenotypes for type 2 diabetes mellitus using various methods such as Random Forest [16], Stepwise Regression [17], Support Vector Regression (SVR) and Genetic Algorithm (GA) [18], Genetic Algorithm (GA) and CatBoost [19], Gradient Boosting [20], and Elastic Net [21]. Stepwise Regression is a model selection method with a gradual selection feature, however, the number of predictor variables included in the model must be selected manually, so that errors in variable selection can occur. In addition, the Stepwise Regression method does not always produce the best model with minimal multicollinearity [22]. The SVR method is a theory adapted from ma-chine learning theory which is intended to solve classification problems but is unable to handle data with a very large number of dimensions [23], [24].

Elastic net has been considered and selected as the preferred method used in this study because it is a method whose number of parameters is not limited by the number of samples as other conventional linear regressions [21]. The elastic net method adapts Ridge regression in predicting and Lasso regression in selecting features [25]. Based on the description above, this study will use the elastic net method to associate SNPs with CKD phenotypes. So far there has been no research related to the topic discussed in this study, so it is not known how the results of research with the elastic net method are. Thus, it is hoped elastic net method can help to associate SNPs with CKD phenotypes with low Mean Absolute Error (MAE) values and good coefficients of determination (R2).

2. Materials and Methods

2.1. Data collection

The candidate gene data for blood albumin amount, which is needed to get SNP data, was taken through the https://www.genecards.org/ website by entering the keyword "blood albumin amount". The candidate gene data were then selected based on the relevance score by only taking the candidate genes that had a relevance score ≥ 25. Then, the candidate gene data were used to query the SNPs in the CGD-MDA1 database which belongs to Yang W et al [26]. The dataset was obtained from the Mouse Phenome Database (MPD) through the website address https://phenome.jax.org/. The phenotype data was derived from the Yuan3 dataset belonging to Sinke A et al. [27] which contained the measurements for blood albumin amount phenotype, which was also accessed through the MPD website. Blood albumin amount is used as a phenotype marker because it is one of the CKD phenotypes [28]. The retrieved SNP data was adjusted to the availability of the strain data in the phenotype database. There were 31 mouse strains used. All datasets were taken from credible websites because there were journals for each used dataset in this study. The Genecards and MPD websites allow users to download datasets in .csv format so they can use them right away.

2.1.1. Data pre-processing

Three processes were carried out at the data pre-processing stage. First, calculate the average of the phenotype values, which is the blood albumin amount, for each strain. The calculation of the average value was carried out because it only needs one representative phenotype value for each strain. Before calculating the average value of each strain, it was standardized first using the StandardScaler function from the Scikit-Learn package. Secondly, fixing the SNP data by deleting SNPs that contain missing value and genotype 'H' can cause confusion because it can be interpreted as base A, base G, base C, or base T. Third, coding the SNP data which represented using binary string based on [29]. SNPs with the most appearing modes are called major alleles, while the rest are called minor alleles. In genotype sequence, allele information is formed by the variations of {A/A, A/T, A/C, A/G ... T/C, T/T}. Based on the conditions of the genotype variation, the SNP encoding into a binary string shown to the following:

0: both alleles are homozygous majors

1: both alleles are homozygous minors

2: both alleles are heterozygous

The example of encoding the SNP data according to Ilhan I et al. can be seen in the following Figure 1. The homozygous major allele is represented by gray, the homozygous minor allele is represented by black, and white is for heterozygous alleles.

	SNP1	SNP2	SNP3	SNP4	SNP5
Individu 1	A/A	T/T	T/T	G/G	A/C
Individu 2	C/G	T/A	C/C	C/G	G/G
Individu 3	A/A	A/T	T/C	C/C	C/A
Individu 4	G/C	T/T	C/C	G/G	G/G

	SNP1	SNP2	SNP3	SNP4	SNP5
Individu 1	0	0	1	0	2
Individu 2	2	2	0	2	0
Individu 3	0	2	2	1	2
Individu 4	2	0	0	0	0

Fig 1. The SNP gene before and after being converted into a binary string.

2.2. Modeling

By using the Grid Search Cross Validation (GridSearchCV), a search was carried out in advance for the best hyperparameters that suit the used model. The most suitable hyperparameters will be used for model training. While the Leave One Out Cross Validation (LOOCV) was used for training and model selection in Figure 2. Then, the method which was used for the machine learning modeling at this stage was Elastic Net by using the Scikit-Learn package which provides a linear model as a module that has an ElasticNet class. This method can directly select features that have significant importance values during modeling. While features that were not significant will be given a regression coefficient value from the model equation = 0, so these features were not selected. Set of selected significant

SNPs that were considered as SNPs that influence the predetermined phenotypes.

2.2.1. Leave One Out Cross Validation (LOOCV)

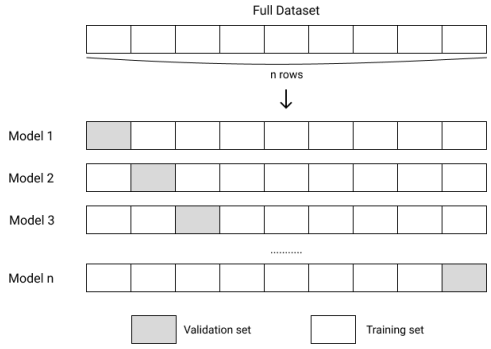

Fig 2. The illustration of how the LOOCV technique works

By using the LOOCV technique, for each training set it is only one data point that is used as test data, and the remaining data will be used as training data for all samples in the data. Therefore, if there are 100 data points, then there are also 100 models for each data point that is used as the test data. This is the LOOCV thing that can overcome the drawbacks of using data with a small number of samples because it does not create partitions between the training and the test data.

2.2.2. Elastic Net method

Elastic Net adapts the feature on Ridge Regression which can reduce the insignificant feature coefficients to close to zero, with features on Lasso Regression which can eliminate insignificant features by making the coefficients equal to zero, thus making Elastic Net able to predict as well as perform feature selection well such in equations (1) and (2).

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_2 \sum_{j=1}^{p}(\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \quad (1)$$

$$\text{With } \alpha = \frac{\lambda_1}{\lambda_1+\lambda_2},\ 0 \leq \alpha \leq 1$$

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda[(1-\alpha)\sum_{j=1}^{p}(\beta_j)^2 + \alpha \sum_{j=1}^{p} |\beta_j|] \quad (2)$$

Where α is the combined parameter. If α = 0 then it will be Ridge Regression function, whereas α = 1 then it will be a Lasso Regression function. λ is a penalty constant with a default value = 1. If λ=0, then the function will be a completely Ordinary Least Square (OLS). β_j is the regression coefficient (slope), and β_0 is a constant (intercept). x is the data value (independent variable), and y is the target value (dependent variable) [30].

2.3. Model testing and evaluation

The testing phase was carried out to predict the quantitative value of the target variable, namely blood albumin amount using the best model that has been obtained in the previous stage when looking for hyperparameters using the LOOCV technique. Then, for the evaluation stage model, using the Mean Absolute Error (MAE) and the coefficient of determination or explained variance (R2). In this study, MAE measured the level of accuracy from the calculation results of the Elastic Net method. While R2 testing aimed to measure how good the regression line is.

2.3.1. Mean Absolute Error (MAE)

If the MAE value is close to zero, then the model is considered good. The following is the equation used to calculate the MAE value [31].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| \quad (3)$$

n = number of data

f_i = the value of the i-prediction results

y_i = the true value

2.3.2. Explained Variance ($\boldsymbol{R^2}$)

Modeling with an R2 value close to 1 is considered good enough, whereas an R2 value close to 0 is considered a bad regression model [32]. The following is an equation for calculating the value (R2) [33].

$$R^2 = 1 - \frac{\sum(y_i - y_c)^2}{\sum(y_i - y_r)^2} \quad (4)$$

y_i = the i-th value

y_c = the estimated value of the regression equation

y_r = the average value

2.4. Validation of SNP selection results

The SNP validation stage that has been selected previously is carried out by mapping the SNP to the gene of origin in the gene data. Then, based on the previous research, a literature study was carried out to see if there were studies that had succeeded in revealing the association between the genes that had been obtained with the phenotype of blood albumin amount or CKD in general as a disease. The literature search used credible article reference sources such as PubMed (https://pubmed.ncbi.nlm.nih.gov/).

3. Result

3.1 Pre-processing data

The used phenotype data was taken from the Yuan3 dataset with a total of 32 different strains, with each strain available in 2 genders, female and male accompanied by the average albumin value for each strain, because this study only took strains aged 12 months, there was one strain that had no

average albumin value at all, MOLF/EiJ, bringing the total to 31 rat strains. Then standardization was carried out with the Scikit-Learn package for these strains so that the distribution value for each strain has the same average and standard deviation. Because only one value is needed to represent each strain, strains that have more than one value will take the average or median value to serve as a representative for the albumin value of each strain. The results were presented in Table 1.

Table 1. Albumin value data for each strain that has been standardized

Strain	The average albumin values which have been standardized
1291/SvImJ	-0.77072
A/J	0.656164
BALB/cByJ	-1.70946
BTBR T<+>Itpr3<tf>/J	-0.188702
BUB/BnJ	-0.977245
C3H/HeJ	0.7688128
C57BL/10J	-0.507874
C57BL/6J	0.186793
C57BLKS/J	0.074144
C57BR/cdJ	1.031660
C57L/J	0.787587
CAST/EiJ	-2.68575
CBA/J	0.280667
DBA/2J	-0.48909
FVB/NJ	-0.75195
KK/HlJ	-1.44662
LP/J	-0.62052
MRL/MpJ	-0.54542
NOD.B10Sn-H2<b>/J	0.900237
NON/ShiLtJ	-1.14622
NZO/HlLtJ	1.425931
NZW/LacJ	0.674939
P/J	-0.9397
PL/J	1.425931
PWD/PhJ	0.750038
RIIIS/J	0.205568
SJL/J	-0.32013
SM/J	1.53858
SWR/J	0.111694
WSB/EiJ	0.149244

Whereas for SNP data pre-processing, improvements were made to handle missing values and 'H' genotypes which could reduce the accuracy level of the model. To overcome this, it is needed to delete rows that still have the genotype value 'H' and the missing values. So, from the 114 candidate genes queried, there are only 77 available gene data in the SNP dataset. The SNP data coding was carried out only by changing the homozygous major allele to 0 value and to the 1 value for the homozygous minor allele. No data was changed to 2 because there were no heterozygous alleles. Then after the value in the SNP data has been changed, the SNP data was combined with the blood albumin amount value data based on the strain. The results of this data pre-processing stage ended in a data frame with dimensions 29 (samples of strains) x 580 (579 SNP features as predictors, 1 target feature which is blood albumin amount) (Table 2).

Table 2. SNP data after coding.

Strain	rs31700936	rs31700937	rs625693	...	rs6161115	The average albumin
1291/SvImJ	0	0	0	…	0	-0.770722
A/J	0	0	0	…	0	0.656164
BALB/cByJ	0	0	0	…	0	-1.709463
…	…	…	…	…	…	…
WSB/EiJ	1	1	0	…	0	0.149244
C57BLKS/J	0	0	0	…	0	0.0741451

3.2 Modeling

Hyperparameter search of the Elastic Net model was performed using the LOOCV-assisted GridSearchCV function from the Scikit-Learn using the search grid configuration shown in the following Table 3.

Table 3. Configuration of alpha and l1_ratio for hyperparameter search.

Alpha	l1_ratio

10^{-5}; 0.0001; 0.001; 0.01; 0.1; 0.0; 1.0; 10.0; 100.0	0; 0.01; 0.02; 0.03; 0.04; 0.05; 0.06; 0.07; 0.08; 0.09; 0.1; 0.11; 0.12; 0.13; 0.14; 0.15; 0.16; 0.17; 0.18; 0.19; 0.2; 0.21; 0.22; 0.23; 0.24; 0.25; 0.26; 0.27; 0.28; 0.29; 0.3; 0.31; 0.32; 0.33; 0.34; 0.35; 0.36; 0.37; 0.38; 0.39; 0.4; 0.41; 0.42; 0.43; 0.44; 0.45; 0.46; 0.47; 0.48; 0.49; 0.5; 0.51; 0.52; 0.53; 0.54; 0.55; 0.56; 0.57; 0.58; 0.59; 0.6; 0.61; 0.62; 0.63; 0.64; 0.65; 0.66; 0.67; 0.68; 0.69; 0.7; 0.71; 0.72; 0.73; 0.74; 0.75; 0.76; 0.77; 0.78; 0.79; 0.8; 0.81; 0.82; 0.83; 0.84; 0.85; 0.86; 0.87; 0.88; 0.89; 0.9; 0.91; 0.92; 0.93; 0.94; 0.95; 0.96; 0.97; 0.98; 0.99

The hyperparameter of the GridSearchCV search results got the best alpha value of 0.01 and an l1_ratio of 0.01. The Elastic Net model with the best hyperparameter succeeded in selecting 88 significant SNPs from the 579 SNPs that had been obtained. Determination of the Elastic Net coefficient obtained from each SNP using a threshold. The selected SNPs are SNPs with a coefficient value of ≥ 0.05 or ≤ -0.05. All SNP coefficient values were converted to absolute value so that what was considered was only the magnitude of the coefficient value and did not pay attention to the direction of the coefficient value. The greater the coefficient value, the more significant the effect of the SNP. The following table will display the 10 SNPs with the highest significance (Table 4).

Table 4. Ten SNPs with the highest value of significance.

SNPid	|Significant value|
rs32017704	0.149330
rs16814674	0.148558
rs50420333	0.126999
rs36885702	0.126723
rs16815124	0.124799
rs38990326	0.117344
rs29886255	0.115257
rs27558506	0.109833
rs30408777	0.108168
rs13483829	0.103778

3.2.1 Evaluation and testing

The test obtained was to make a comparison between the predicted value and the actual value of the blood albumin amount using SNP as the predictor variable. For comparison can be seen in the following Table 5.

Table 5. The test results contain comparisons.

Strain	Actual value	Predicted value
129S1/SvImJ	-0.770722	-0.752759
A/J	0.656164	0.646480
BALB/cByJ	-1.709463	-1.693074
…	…	…
WSB/EiJ	0.149244	0.145055
C57BLKS/J	0.074145	0.073743

Then the Elastic Net model in this study was evaluated using the MAE and R2 metrics. The MAE results show a value of 0.755 which means a low failure rate. While the R2 obtained was 0.999 or almost close to 1 indicating that the variation in the value of blood albumin amount can be explained by the SNP as a predictor that has been selected by the model.

4. Discussion

The Elastic Net model selects SNPs along with the modeling process. The selected significant SNPs were 88 out of 579 processed SNPs. The 88 SNPs represent 30 candidate genes out of 77 previously selected candidate genes for querying. Furthermore, by conducting a literature study, it was found that as many as 19 out of 30 (63.33%) gene variants previously obtained, were shown to have associations with CKD or other kidney diseases as well as phenotype identifiers for blood albumin amount. While 11 other gene types have not been found in published literature studies related to CKD or blood albumin amount phenotypes. A complete description of the genes and their relation to CKD can be seen in the following table 6.

Table 6. Validation of SNP selection based on literature study.

Gene type	Relation with CKD
F2	Associated with kidney stone disease in women in Thailand, is one of the phenotypes of CKD (Rungroj N, Sudtachat N, Nettuwakul C et al. [34])
Col1a1	This gene is associated with renal allograft fibrosis which is one of the CKD phenotypes (Menom M, Chuang P, Li Z et al. [35])
Vfw	This gene is a predictor of microalbuminuria, which is one of the CKD phenotypes (Perrson F, Rossing P, Hovind P et al. [36])

Nos3	Glu298Asp polymorphisms 4 b/a and -786>C of Nos3 were shown to have an association with CKD (Medina A, Zubero E, Jim´enez M et al. [37])
Cd44	The Cd44 SNP polymorphism is associated with kidney stone disease, which is one of the symptoms of CKD, taking samples from the Han Tinghoa population in Northeast Sichuan, China (Ying Q, Liu G, Zhou W et al. [38])
F10	SNP gene F10 has been shown to be associated with one of the symptoms of CKD (Yoshida T, Kato K, Yokoi K et al. [39])
Bcam	The Lu/Bcam protein plays a role in presenting the accumulation of monocytes and macrophages thereby exacerbating kidney injury which can refer to CKD (Huang J, Filipe A, Rahuel C et al. [40])
Runx1	RNA sequencing revealed that runx1-runx1t1 can cause Clear Cell Renal Cell Carnicoma (ccRCC) which can lead to kidney cancer (Xiong Z, Yu H, Ding Y et al. [41])
Itgb3	Itgb3 plays a role in tubular cell aging which can refer to CKD (Li S, Jiang S, Zhang Q et al. [42])
Cfh	This gene is found in pregnant women who suffer from hemolytic uremic syndrome which refers to Acute Kidney Injury (AKI) which if it becomes severe will become CKD (Bruel A, Kavanagh D, Noris M et al. [43])
Rbp4	A study was conducted on children who had CKD with excess A intake and found a high number of Rbp4 genes (Manichavasagar B, McArdle A, Yadav P et al. [44])
Flt1	Plasma sFlt-1 levels are inversely correlated with eGFR and are directly related to heart failure in CKD patients (Di Marco G, Kentrup D, Reuter S et al. [45])
Plg	Patients who wish to do hemodialysis must first check their levels of Plg, TC, TG, HDL-C, LDL-C, TC/HDL-C ratio, lipoprotein A, FG, FN, and HCT (Tzanatos H, Tseke P, Pipili C et al. [46])
Myh9	Myh9 mutations were detected in boys with hematuria and proteinuria suggestive of CKD (Matsumoto T, Yanagihara T, Yoshizaki K et al. [47])
ApoB	ApoB monitoring can be helpful for prediction of End Stage Renal Disease (ESRD) (Kwon S, Kim D, Oh K et al. [48])
Eng	Heterogeneous expression of endoglin characterizes advanced kidney cancer with different tumor microenvironmental states (Momoi Y, Nishida J, Miyakuni K et al. [49])
Lep	Increasing the number of these genes in CKD patients can worsen kidney function and lead to increased cardiovascular risk (Korczynska J, Czumaj A, Chmielewski M et al. [50])
Fgb	Indexed chronic glomerulonephritis patient had the FGB G455A polymorphism (Kozlovskaia, N L et al. [51])
Akt1	This gene is thought to help fight CKD (Lin H, Chen Y, Chen Y et al. [52])

5. Conclusions

This study succeeded in associating SNP as a predictor feature with the blood albumin amount phenotype in CKD using a machine learning method, namely Elastic Net Regression. The Elastic Net model used produces the best hyperparameters for alpha of 0.01 and l1 ratio of 0.01. The evaluation results for the Elastic Net model in this study also show a low error rate, the MAE is 0.755, and the coefficient of determination is 0.999. For the selected SNPs, there were 88 SNPs in this study, with 19 of the 30 genes selected through a literature study. The highest significance value is SNPid rs32017704 chromosome 5 of 0.149330 which is the SNP of the Nos3 gene, and it is proven from literature studies that the Nos3 gene is related to CKD. Followed by SNPid rs16814674 from the F7 gene, SNPid rs50420333 from the Plg gene, rs36885702 from the Fgb gene, and so on.

Acknowledgements

The authors would like to thank Multimedia Nusantara University for their support of this research work.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] R. A. Wijayanti, M. T. Furqon, and S. Adinugroho, "Penerapan Algoritme Support Vector Machine Terhadap Klasifikasi Tingkat Risiko Pasien Gagal Ginjal," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[2] Yulianti, R. Amegia Saputra, M. Sukrisno Mardiyanto, and A. Rahmawati, "Optimasi Akurasi Algoritma C4.5 Berbasis Particle Swarm Optimization dengan Teknik Bagging pada Prediksi Penyakit Ginjal Kronis Optimization of C4.5 Algorithm Based On Particle Swarm Optimization with Bagging Technique on Prediction of Chronic Kidney Disease," 2020. [Online]. Available: https://archive.ics.uci.edu/ml/

[3] O. : Anggun, H. Safitri,) Dinar, S. E. Dewi, and) Abstrak, "DESKRIPSI TINGKAT HARAPAN PADA PENDERITA GAGAL GINJAL KRONIK DI RSU PROF DR. MARGONO SOEKARJO PURWOKERTO DESCRIPTION OF HOPE IN CHRONIC RENAL FAILURE PATIENTS IN RSU PROF DR. MARGONO SOEKARJO PURWOKERTO."

[4] "InfoDATIN," 2006. [Online]. Available: http://emojione.com

[5] U. N. Semarang, M. Masa, D. Pendidikan, I. Hayati, P. Pandemi, and T. N. Azhar1, "Prosiding Seminar Nasional Biologi X FMIPA."

[6] B. Li et al., "GPCards: An integrated database of genotype–phenotype correlations in human genetic diseases," Comput Struct Biotechnol J, vol. 19, pp. 1603–1611, Jan. 2021, doi: 10.1016/j.csbj.2021.03.011.

[7] Meiliana, N. M. Dewi, and A. Wijaya, "Metabolomics: An Emerging Tool for Precision Medicine," Indonesian Biomedical Journal, vol. 13, no. 1, pp. 1–18, 2021, doi: 10.18585/inabj.v13i1.1309.

[8] R. L. Perlman, "Mouse Models of Human Disease: An Evolutionary Perspective," Evol Med Public Health, p. eow014, Apr. 2016, doi: 10.1093/emph/eow014.

[9] J. P. Ly, T. Onay, and S. E. Quaggin, "Mouse models to study kidney development, function and disease," Current Opinion in Nephrology and Hypertension, vol. 20, no. 4. pp. 382–390, Jul. 2011. doi: 10.1097/MNH.0b013e328347cd4a.

[10] G. Research, "Help Me Understand Genetics." [Online]. Available: https://medlineplus.gov/genetics/

[11] Y. Dwiningsih, M. Rahmaningsih, and J. Alkahtani, "Development of Single Nucleotide Polymorphism (SNP) Markers in Tropical Crops," Advance Sustainable Science, Engineering and Technology, vol. 2, no. 2, Jul. 2020, doi: 10.26877/asset.v2i2.6279.

[12] H. Suprapti, B. Farmakologi, F. Kedokteran, U. Wijaya, and K. Surabaya, "Farmakogenomik Statin: Biomarker untuk Prediksi Klinis," Online) Jurnal Ilmiah Kedokteran Wijaya Kusuma, vol. 7, no. 1. pp. 1–14, 2018.

[13] T. Marees et al., "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," Int J Methods Psychiatr Res, vol. 27, no. 2, Jul. 2018, doi: 10.1002/mpr.1608.

[14] T. Nguyen and L. Le, "Detection of SNP-SNP Interactions in Genome-wide Association Data Using Random Forests and Association Rules," in 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), 2018, pp. 1–7. doi: 10.1109/SKIMA.2018.8631529.

[15] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification," Journal of Computer Science, vol. 14, no. 11, pp. 1521–1530, 2018, doi: 10.3844/jcssp.2018.1521.1530.

[16] L. H. Tresnawati, W. A. Kusuma, S. H. Wijaya, and L. S. Hasibuan, " Asosiasi Single Nucleotide Polymorphism pada Diabetes Mellitus Tipe 2 Menggunakan Random Forest Regression," Jurnal Nasional Teknik Elektro dan Teknologi Informasi, vol. 8, no. 4, pp. 320–326, Nov. 2019, [Online]. Available: https://jurnal.ugm.ac.id/v3/JNTETI/article/view/2556

[17] "ASOSIASI SINGLE NUCLEOTIDE POLYMORPHISM DAN FENOTIPE PADA PENYAKIT DIABETES MELLITUS TIPE 2 MENGGUNAKAN STEPWISE REGRESSION DEVY APRIANSYAH."

[18] R. M. Siregar, "Asosiasi single nucleotide polymorphism pada penyakit diabetes mellitus tipe 2 menggunakan support vector regression dan genetic algorithm," IPB University.

[19] H. F. Ramadhani, W. A. Kusuma, L. S. Hasibuan, and R. Heryanto, "Association of single nucleotide polymorphism and phenotypes in type 2 diabetes mellitus using genetic algorithm and catboost," in 2020 International Conference on Computer Science and Its Application in Agriculture, ICOSICA 2020, Institute of Electrical and Electronics Engineers Inc.,

Jul. 2020. doi: 10.1109/ICOSICA49951.2020.9243208.

[20] Fadli, L. S. Hasibuan, W. A. Kusuma, and R. Heryanto, "Single nucleotide polymorphism and type 2 diabetes mellitus phenotypes association using gradient boosting," in 2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020, Institute of Electrical and Electronics Engineers Inc., Jul. 2020, pp. 115–120. doi: 10.1109/ICACSIS51025.2020.9263142.

[21] M. F. Romdendine, "ASOSIASI SINGLE NUCLEOTIDE POLYMORPHISM DAN FENOTIPE PADA PENYAKIT DIABETES MELLITUS TIPE 2 MENGGUNAKAN METODE ELASTIC NET," 2022.

[22] H. Hanum, "Perbandingan Metode Stepwise, Best Subset Regression, dan Fraksi dalam Pemilihan Model Regresi Berganda Terbaik," Jurnal Penelitian Sains, vol. 14. p. 14201.

[23] N. D. Maulana, B. D. Setiawan, and C. Dewi, "Implementasi Metode Support Vector Regression (SVR) Dalam Peramalan Penjualan Roti (Studi Kasus: Harum Bakery)," vol. 3, no. 3. pp. 2986–2995, 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[24] M. Tingkatkemanisan, M. Berdasarkan, and F. Warna, "Klasifikasi Support Vector Machine (SVM) Untuk," MIND Journal | ISSN, vol. 3, no. 2, pp. 16–24, 2018, doi: 10.26760/mindjournal.

[25] Comber and P. Harris, "Geographically weighted elastic net logistic regression," J Geogr Syst, vol. 20, no. 4, pp. 317–341, Jul. 2018, doi: 10.1007/s10109-018-0280-7.

[26] H. Yang et al., "Subspecific origin and haplotype diversity in the laboratory mouse," in Nature Genetics, Jul. 2011, pp. 648–655. doi: 10.1038/ng.847.

[27] P. Sinke et al., "Genetic analysis of mouse strains with variable serum sodium concentrations identifies the Nalcn sodium channel as a novel player in osmoregulation," Physiol Genomics, vol. 43, pp. 265–270, 2011, doi: 10.1152/physiolgenomics.00188.2010.-In.

[28] J. Van De Wouw and J. A. Joles, "Albumin is an interface between blood plasma and cell membrane, and not just a sponge," Clinical Kidney Journal, vol. 15, no. 4. Oxford University Press, pp. 624–634, Apr. 01, 2022. doi: 10.1093/ckj/sfab194.

[29] Ilhan and G. Tezel, "How to select tag SNPs in genetic association studies? the CLONTagger method with parameter optimization," OMICS, vol. 17, no. 7, pp. 368–383, Jul. 2013, doi: 10.1089/omi.2012.0100.

[30] H. Zou and T. Hastie, "Erratum: Regularization and variable selection via the elastic net (Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320))," J R Stat Soc Series B Stat Methodol, vol. 67, no. 5, p. 768, 2005, doi: 10.1111/j.1467-9868.2005.00527.x.

[31] Technology University of Oradea. Faculty of Electrical Engineering and Information, IEEE Romania Section. CAS/CA Chapter, B. Association of Romanian Electrical and Electronics Engineers, O. Association of Integrated Engineering and Industrial Management, and Institute of Electrical and Electronics Engineers, 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) : Oradea, România, June 01-02, 2017.

[32] Sukmana, "Koefisien Determinasi R^2 pada Model Regresi Linear," 1996.

[33] C. Cameron and F. A. G. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models."

[34] N. Rungroj et al., "Association between Human Prothrombin Variant (T165M) and Kidney Stone Disease," PLoS One, vol. 7, no. 9, Sep. 2012, doi: 10.1371/journal.pone.0045533.

[35] M. C. Menon et al., "Intronic locus determines SHROOM3 expression and potentiates renal allograft fibrosis," Journal of Clinical Investigation, vol. 125, no. 1, pp. 208–221, Jan. 2015, doi: 10.1172/JCI76902.

[36] F. Persson et al., "Endothelial dysfunction and inflammation predict development of diabetic nephropathy in the Irbesartan in Patients with Type 2 Diabetes and Microalbuminuria (IRMA 2) study," Scand J Clin Lab Invest, vol. 68, no. 8, pp. 731–738, Dec. 2008, doi: 10.1080/00365510802187226.

[37] M. Medina et al., "NOS3 Polymorphisms and Chronic Kidney Disease," Jornal brasileiro de nefrologia : 'orgao oficial de Sociedades Brasileira e Latino-Americana de Nefrologia, vol. 40, no. 3. NLM (Medline), pp. 273–277, Jul. 01, 2018. doi: 10.1590/2175-8239-JBN-3824.

[38] Q. Ying et al., "The rs13347 Polymorphism of the CD44 Gene Is Associated with the Risk of Kidney Stones Disease in the Chinese Han Population of Northeast Sichuan, China," Comput Math Methods Med, vol. 2022, 2022, doi: 10.1155/2022/6481260.

[39] polymorphisms with chronic kidney disease in Japanese individuals," Int J Mol Med, vol. 24, no. 4, pp. 539–547, 2009, doi: 10.3892/ijmm_00000263.

[40] J. Huang et al., "Lutheran/basal cell adhesion molecule accelerates progression of crescentic glomerulonephritis in mice," Kidney Int, vol. 85, no. 5, pp. 1123–1136, 2014, doi: 10.1038/ki.2013.522.

[41] Z. Xiong et al., "RNA sequencing reveals upregulation of RUNX1-RUNX1T1 gene signatures in clear cell renal cell carcinoma," Biomed Res Int, vol. 2014, 2014, doi: 10.1155/2014/450621.

[42] S. Li et al., "Integrin β3 Induction Promotes Tubular Cell Senescence and Kidney Fibrosis," Front Cell Dev Biol, vol. 9, Nov. 2021, doi: 10.3389/fcell.2021.733831.

[43] Bruel et al., "Hemolytic uremic syndrome in pregnancy and postpartum," Clinical Journal of the American Society of Nephrology, vol. 12, no. 8, pp. 1237–1247, 2017, doi: 10.2215/CJN.00280117.

[44] [B. Manickavasagar et al., "Hypervitaminosis A is prevalent in children with CKD and contributes to hypercalcemia," Pediatric Nephrology, vol. 30, no. 2, pp. 317–325, Aug. 2014, doi: 10.1007/s00467-014-2916-2.

[45] G. S. Di Marco et al., "Soluble Flt-1 links microvascular disease with heart failure in CKD," Basic Res Cardiol, vol. 110, no. 3, Apr. 2015, doi: 10.1007/s00395-015-0487-4.

[46] H. A. Tzanatos, P. P. Tseke, C. Pipili, K. Retsa, G. Skoutelis, and E. Grapsa, "Cardiovascular risk factors in non-diabetic hemodialysis patients: A comparative study," Ren Fail, vol. 31, no. 2, pp. 91–97, Feb. 2009, doi: 10.1080/08860220802595484.

[47] T. Matsumoto et al., "Renal Biopsy-induced Hematoma and Infection in a Patient with Asymptomatic May-Hegglin Anomaly," Journal of Nippon Medical School, vol. 88, no. 6, pp. 579–584, 2021, doi: 10.1272/JNMS.JNMS.2021_88-609.

[48] S. Kwon et al., "Apolipoprotein B is a risk factor for end-stage renal disease," Clin Kidney J, vol. 14, no. 2, pp. 617–623, Feb. 2021, doi: 10.1093/ckj/sfz186.

[49] Y. Momoi et al., "Heterogenous expression of endoglin marks advanced renal cancer with distinct tumor microenvironment fitness," Cancer Sci, vol. 112, no. 8, pp. 3136–3149, Aug. 2021, doi: 10.1111/cas.15007.

[50] J. Korczynska, A. Czumaj, M. Chmielewski, J. Swierczynski, and T. Sledzinski, "The causes and potential injurious effects of elevated serum leptin levels in chronic kidney disease patients," International Journal of Molecular Sciences, vol. 22, no. 9. MDPI, May 01, 2021. doi: 10.3390/ijms22094685.

[51] N. L. Kozlovskaya et al., "Clinicomorphological characteristics of renal disorders in patients with genetic thrombophilia," Ter Arkh, vol. 81, no. 8, pp. 30–36, 2009, [Online]. Available: https://ter-arkhiv.ru/0040-3660/article/view/30483

[52] H. Y. H. Lin et al., "Tubular mitochondrial AKT1 is activated during ischemia reperfusion injury and has a critical role in predisposition to chronic kidney disease," Kidney Int, vol. 99, no. 4, pp. 870–884, Apr. 2021, doi: 10.1016/j.kint.2020.10.038.