ILKOM Jurnal Ilmiah Vol. 14, No. 3, December 2022, pp. 194-202
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Association of single nucleotide polymorphism and phenotype in type 2 of diabetes mellitus using Support Vector Regression and Genetic Algorithm

Ratu Mutiara Siregar [a,1,*]; Wisnu Ananta Kusuma [a,2]; Annisa[a,3]
[a] *IPB University, Jl. Raya Dramaga, Bogor and 16680, Indonesia*
[1]*ratu_ms@apps.ipb.ac.id,* [2]*ananta@apps.ipb.ac.id,* [3]*annisa@apps.ipb.ac.id*
** Corresponding author*

Article history: Received June 15, 2022; Revised August 20, 2022; Accepted November 15, 2022; Available online December 20, 2022

Abstract

Precision Medicine is used to improve proper health care and patients' quality of life, one of which is diabetes. Diabetes Mellitus (DM) is a multifactorial and heterogeneous group of disorders characterized by deficiency or failure to maintain normal glucose homeostasis. About 90% of all DM patients are Type 2 Diabetes Mellitus (T2DM). Biological characteristics and genetic information of T2DM disease were obtained by looking for associations in Single Nucleotide Polymorphism (SNP) which allows for determining the relationship between phenotypic and genotypic information and identifying genes associated with T2DM disease. This research focuses on the Support Vector Regression method and Genetic Algorithm to obtain SNPs that have previously calculated the correlation value using Spearman's rank correlation. Then do association mapping on the SNP results from the SVR-GA selection and check pastasis interaction. The results produced 14 SNP importance. Evaluation of the model using the mean absolute error (MAE) obtained is 0.02807. If the value of MAE is close to zero, then a model can be accepted. The genes generated from the association can be used to assist other researchers in finding the right treatment for T2DM patients according to their genetic profile.

Keywords: Aassociation mapping; Epistasis; SNP; SVR-GA; Type 2 Diabetes mellitus

Introduction

Precision Medicine is a modern concept that has been used to accurately describe medical treatments tailored to the individual characteristics of each patient. Precision Medicine is used to improve good health services and patients' quality of life, one of which is Diabetes [1]. The current classification of Diabetes into two main forms, Diabetes Mellitus Type 1 (T1DM) or called insulin-dependent Diabetes Mellitus, and Diabetes Mellitus type 2 (T2DM) or called non-insulin-dependent Diabetes Mellitus [1]. T2DM is a diagnosis of exclusion, where if the patient does not have T1D or Monogenic Diabetes, he is considered to have T2DM. This suggests that about 90% of all patients with Diabetes will develop T2DM [2].

Research on human drug and disease testing covers the fields of biomedical, social, epidemiological, and behavioral. Therefore, drug testing can be carried out using experimental animal intermediaries. Animals that are often used as intermediaries for drug testing are mice because 95% of mouse genes resemble humans [3].

A person's biological state can be characterized by genomics data types consisting of genotypes. The genotype has markers related to the phenotype in T2DM disease. It is called single nucleotide polymorphism (SNP-read snip) [4]. Phenotype is an observable trait, such as height, skin pigmentation, and response to a drug [5]. The number of complex human diseases is thought to be not only caused by one genetic variation but is determined by epistasis, namely the interaction between many genes [6]. Associations in SNPs allow for determining the relationship between genotypic and phenotypic information and identifying disease-associated genes [7]. The association of SNPs with phenotypes in T2DM can assist in precise precision medicine and yield useful information for drug discovery used to treat T2DM disease.

http://dx.doi.org/10.33096/ilkom.v14i3.1283.194-202

Previous research related to the association method on SNP, including the research of Ban [8], used the Support Vector Machine to analyze the interaction genes of 408 SNPs and 87 genes associated with T2DM disease in the Korean population. Ilhan and Tezel [9] conducted a selection of genetic variations to explain individual differences in complex diseases using the same method, namely the Support Vector Machine (SVM). Research [10] also conducted the same study using the Random Forest method to rank SNPs. The ranking results were used to perform regression and produce 301 SNPs associated with the T2DM disease phenotype, where the number is still relatively large for laboratory testing and increases the cost of testing.

Therefore, this study will propose the Support Vector Regression and Genetic Algorithm methods to select significant SNPs by first calculating the correlation between SNPs and insulin tolerance as a phenotype using Spearman correlation. After that, association mapping and evaluation of epistasis or the interaction between the two associated SNPs were carried out. The results of this SNP-phenotype association study are expected to be able to carry out a precision medicine approach to get the proper treatment according to the genetic profile of each T2DM patient.

Method

There are several stages carried out in this research. The stage begins with collecting SNP data and then preprocessing the data. After cleaning the data, SNP correlation analysis on insulin tolerance was performed. Variables that correlate with 0.20 will proceed to the next stage, namely feature selection. The best SNP candidates were selected using Support Vector Regression with a Genetic Algorithm (SVR-GA). The SVR-GA model with the best features was used to analyze the association mapping of SNPs for T2DM-related phenotypes. The stages of the research can be seen in **Figure 1**.

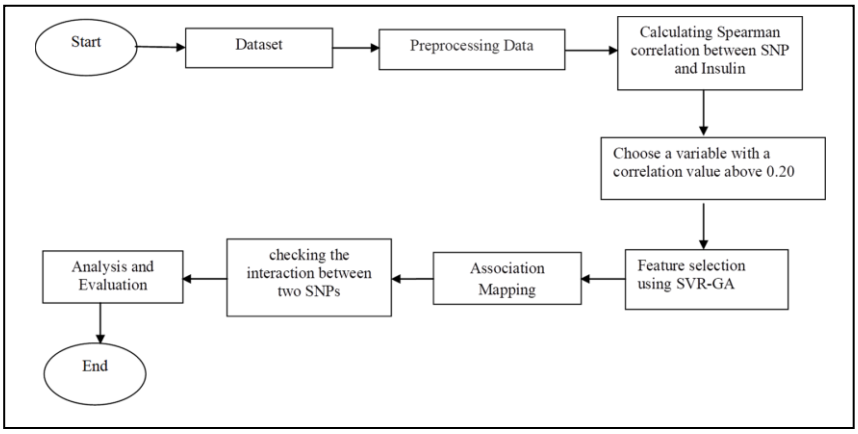

Figure 1. Research Stages

A. Data Collection

The data used is a T2DM disease SNP dataset from mouse genes found on the Mouse Phenome Database (MPD) site with the website address https://phenome.jax.org/snp. The SNP query search was based on 71 candidate proteins associated with the T2DM gene. The phenotype data used were insulin tolerance from the CGD-MDA1 database of as many as 9 male rat strains. CGD-MDA1 is a 2014 inbred mouse database with more than 470,000 genomic locations.

Data Preprocessing

SNP data preprocessing was done by merging the downloaded SNP query results from each gene on the Mouse Phenome Database site with the website address https://phenome.jax.org/snp. After that, eliminate the unused variables because the variables used are only strains and SNPs, then eliminate the missing value. In the phenotype data, preprocessing is done by removing features or variables that are not used because the only variable used is the insulin tolerance value of 9 strains. Then it is also because each mice has an insulin tolerance value of ten. The insulin tolerance data is calculated using the mean.

The allele used in this study is the biallelic allele because mice are diploid animals. Next, the SNP encoding is performed using the scrime package on Rstudio version 3.6.0 and the SNP recode function. Coding for biallelic SNPs is represented in three numbers [11]. SNP coding was performed based on allele information on the genotype sequences formed by variations {A/A, A/T, A/G, A/C…G/T, G/C}. The coding of SNPs into three numbers based on the conditions of genotype variation is as follows:

1: both alleles are major homozygous,

2: both alleles are heterozygous,

3: both alleles are minor homozygous.

B. Calculating the Correlation between SNP and Insulin Tolerance

In this study, the coefficient of the relationship between SNP and insulin tolerance was calculated to see the level of correlation. It is useful to see which variables have a true correlation with insulin tolerance. SNP data is a numerical variable, while the dependent variable is continuous. Thus, the correlation calculation method used is Spearman. The limit of the coefficient value taken in this study is 0.20, where 0.20 is still a value that can be considered even though the correlation level is weak [12]. The value of the Spearman correlation level can be seen in **Table 1**.

Table 1. Spearman correlation value and degree of relationship

Correlation Value(rs)	Measure of the strength
0.01-0.19	Negligible
0.20-0.39	Weak
0.40-0.59	Moderate
0.60-0.79	Strong
0.80-1.00	Very strong

Spearman correlation is suitable for data with small and discrete samples. The Spearman correlation takes advantage of the difference in rank between two variables. The calculation of the Spearman Correlation can be seen in **Equation 1**.

$$r_s = \frac{1-6\sum {d_i}^2}{n(n^2-1)} \quad (1)$$

Where d_i describes difference between the two observations of SNP and iinsulin tolerance, n describes total observations.

C. SNP Selection using SVR-GA

Implementation GA into the SVR, this study uses variable bits in binary form as individuals. If the variable bit is 1, the variable will be used in the SVR, otherwise the bit will be 0. After that, the bits will be converted into variables that will be used in building the SVR model. The performance of the model will be evaluated for each individual with each SNP combination using MAE.

The results of this evaluation will be the fitness function of GA. After evaluation, the following steps, such as selection, crossover, and mutation, are carried out to update the individual feature combinations. This step is carried out until the iteration is complete. The steps taken in running SVR-GA can be seen in **Figure 2**.

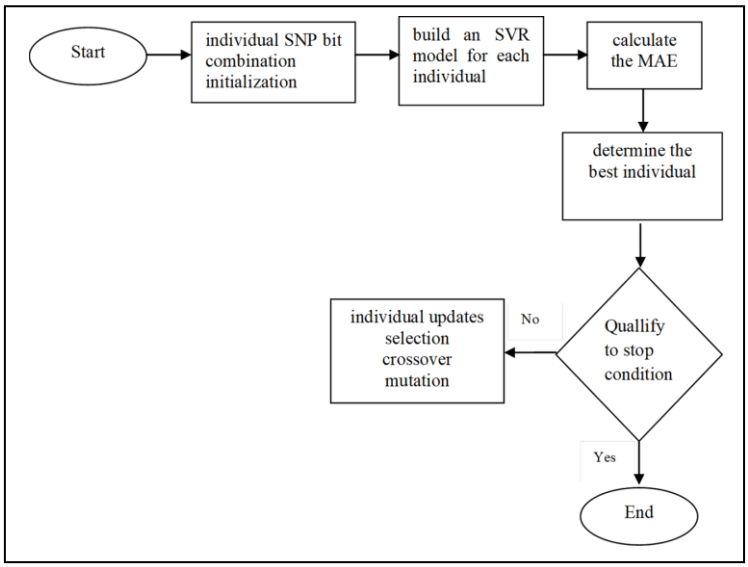

Figure 2. SVR-GA selection for association

D. Association Mapping

After calculating the importance score, association mapping is done after getting the optimal SVR model. The importance score is obtained by finding the SVR coefficient. The SVR coefficient is obtained by multiplying the support vector by the hyperplane formed. However, this can only be done using a linear kernel by multiplying the matrix between the support vector coefficients. As a result, after selecting the SVR kernel RBF, a remodel was made with a linear kernel to get the importance score.

SNP and phenotype associations are analyzed by mapping SNPs used as features in the SVR model with SNPid genetic information, chromosomes, alleles, and genes. The selected SNPs were considered associated with the phenotype in T2DM disease and will then be validated through other literature studies.

E. Checked Epistasis Interaction

SNP interaction was carried out by examining the selection results produced by SVR-GA. The p-value represents the epistasis parameter. If the p-value is 0.1 and positive, then the epistasis between SNPs is in the high category. If the p-value is 0.1 and negative, then the epistasis between SNPs is in the low category, and SNP epistasis is in the undetermined category if the p-value is > 0.1 [13].

F. Evaluation of Association Results

The evaluation of the model uses the mean absolute error (MAE) which represents the error value of the model, which is found in **Equation 2** where if the MAE value is close to zero, then the model is getting better [14].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \qquad (2)$$

n is a amount of data, f_i is predictive value to i, and y_i is a true value.

Results and Discussion

A. Data Collection

The SNP data collection was retrieved from The Mouse Phenome Database (MPD) website at https://phenome.jax.org/snp/retrieval. MPD is an integrated resource for exploring physiology and behavior through genetics and genomics.

Data were contributed by researchers from around the world and represent a wide range of behavioral, morphological, and physiological characteristics associated with disease in those exposed to drugs or other treatments. MPD provides a venue for complying with data-sharing policies and facilitates data reuse and integration to analyze trait relationships, discover the biological basis of complex traits, and assess replication and reproducibility across experimental conditions and protocols. A summary of data obtained from the MPD website can be seen in **Table 2**.

Table 2. Summary Data dari situs MPD

Data component	Description
genotype	
update	17 January 2022
Strain chromosome	9 mice strain 1 – 19, X
number of alleles	Adenin (A) = 4754 Timin (T)= 4586 Guanin (G) = 5669 Sitosin (C)= 5521
total of all SNPs	2053
phenotype	
update	19 January 2022
insulin tolerance	90 values

B. Data Preprocessing

The preprocessing performed on the SNP data combines the downloaded SNP data separately based on the protein candidate query. The variables used were only SNP names and nine mouse strains. After that, data exploration was carried out, apparently missing values and H nucleotide bases. H nucleotide bases are A/T/C nucleotides. There were 17 missing values and eight nucleotide bases. Because the missing values were few and the H nucleotide base could not be determined to be A/T/C, the data was removed. The data were originally from 2053 SNPs and nine strains to 2028 SNPs and nine mouse strains. The data is converted into biallelic form then the SNP can be encoded. Some of the data in biallelic form can be seen in **Figure 3.**

SNP	129X1/SvJ	BALB/cByJ	C3H/HeJ	C57BL/6J	DBA/1LacJ	NZB/BINJ	PL/J	SWR/J	C57BL/6ByJ
rs32028999	AA	GG	GG	GG	GG	GG	GG	GG	GG
rs32765530	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs31650652	TT	CC	CC	CC	CC	CC	CC	CC	CC
rs32759970	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs32763140	CC	TT	TT	TT	TT	TT	TT	TT	TT
rs32764612	TT	CC	CC	CC	CC	CC	CC	CC	CC
rs32767993	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs32769537	CC	CC	CC	CC	CC	CC	CC	CC	CC

Figure 3. Some of the data is in biallelic form

SNP encoding uses the recodeSNPs function in the scrime package on Rstudio. After that, the data is transposed, so that the SNP becomes a variable. Some of the results of SNP coding can be seen in **Figure 4**.

strain	rs3202899	rs3276553	rs3165065	rs3275997	rs3276314
129X1/SvJ	3	1	3	1	3
BALB/cByJ	1	1	1	1	1
C3H/HeJ	1	1	1	1	1
C57BL/6J	1	1	1	1	1
DBA/1LacJ	1	1	1	1	1
NZB/BINJ	1	1	1	1	1
PL/J	1	1	1	1	1
SWR/J	1	1	1	1	1
C57BL/6ByJ	1	1	1	1	1

Figure 4. Partial SNP encoding results

The insulin tolerance data was calculated using the average. The average insulin tolerance obtained can be seen in **Table 3**.

Table 3. Mean value of insulin tolerance in strains

No	Strain	*Insulin tolerance*
1	129X1/SvJ	0.059
2	BALB/cByJ	0.746
3	C3H/HeJ	0.303
4	C57BL/6J	0.451
5	DBA/1LacJ	0.485
6	NZB/BlNJ	0.033
7	PL/J	1.165
8	SWR/J	0.202
9	C57BL/6ByJ	0.414

C. Calculating the Correlation between SNP and Insulin Tolerance

As an initial step, the selection of variables using correlation is carried out from the results of the Spearman correlation between SNP and insulin tolerance. There are 1160 whose correlation cannot be calculated. This happens because the SNP data has the same input (all of 1 or all of 3), which causes the correlation cannot be calculated.

After that, the correlation is ordered based on the absolute value of the correlation between these variables. Variables that correlate more than 0.2 will continue to the next stage. Therefore, the results of this selection obtained 158 variables that correlate with more than 0.20.

D. SNP Selection using SVR-GA

It is necessary to form individuals where the number of SNPs is 158 SNPs to conduct the selection. So, in 1 individual, there are 158 SNP bits in binary form. The population is determined using a small population of 50 individuals. The result sought is the maximum fitness value. In other words, if the fitness value is reversed, the MAE value will approach zero. At this stage, the search stops at the 487th iteration because there is no change in the best fitness value for 50 iterations. Thus, 35 SNPs were obtained from the selection using GA. From the results of the SVR model using a linear kernel, the coefficient matrix obtained from the simulation results is a = [0.0288 −0.0245 0.1087 0.0136 −0.0931 0.0167 −0.0502].

Meanwhile, using the importance score value, the best SNP is obtained by multiplying the matrix between the coefficients and the support vector formed from the model with a linear kernel. The results of the selection of 35 SNPs produced 14 SNPs with the best importance score in this study, which can be seen in **Table 4.**

Table 4. 14 SNPs with the best importance score

SNPid	Importance score
rs30188721	0.25078
rs6312932	0.11922
rs33562020	0.09105
rs36330341	0.09037
rs32178367	0.08644
rs33062250	0.06876
rs33204971	0.06534
rs33254574	0.06064
rs8277066	0.05014
rs28149001	0.04209
rs46846246	0.03086
rs48535529	0.00930
rs33144203	0.00883
rs32028999	0.00335

E. Association Mapping

There were 35 SNPs obtained from the SVR-GA selection, but 21 SNPs were deleted because they had the same SNP information. Thus the number of SNPs used in association mapping is 14 SNPs. **Table 5** shows the results of the association between SNPs and insulin tolerance along with candidate genes.

Table 5. The result of association of SNP to insulin tolerance

SNP id	Chr	Observed Allele	Gen
rs30188721	9	A/G	Lipc
rs6312932	5	A/G	Caln1
rs33562020	5	A/G	Mphosph9
rs36330341	8	G/T	Fto
rs32178367	7	A/G	Akt2
rs33062250	5	C/G	Caln1
rs33204971	17	C/G	Vit
rs33254574	5	C/T	Caln1

SNP id	Chr	Observed Allele	Gen
rs8277066	1	A/G	Capn10
rs28149001	4	C/T	Faf1
rs46846246	12	C/T	Tpo
rs48535529	8	A/C	Insr
rs33144203	17	A/G	Itpr3
rs32028999	7	A/G	Abcc8

F. Checked Epistasis Interaction

After the results of the association, an analysis of the interaction between SNPs was carried out. Before starting the interaction analysis, it is essential to see which variables affect insulin tolerance. This check can be done by using multiple linear regression. After obtaining 14 SNPs that produce associations, 7 SNPs can be calculated for their contribution to insulin tolerance. Epistasis values for one SNP can be seen in **Table 6**.

Table 6. Epistasis value for one SNP

SNPid	p-value	Epistasis	Gen
rs30188721	0.043	High	LIPC
rs32178367	0.040	High	AKT2
rs8277066	0.037	High	CAPN10
rs36330341	-0.052	Low	FTO
rs28149001	0.129	*undetermined*	FAF1
rs48535529	-0.083	Low	INSR
rs32028999	0.349	*undetermined*	ABCC8

With a significance level of 10%, it can be seen that 2 SNPs do not significantly affect insulin, including rs28149001 and rs32028999. In addition, 5 SNPs had a significant effect on insulin tolerance. The significant variables can be used to find interactions between variables. The interaction results of the two SNPs can be seen in **Table 7**.

Table 7. The result of the interaction of two SNP

SNPid	p-value	Epistasis	Gen
rs36330341 : rs48535529	0.0545	High	FTO : INSR

Of the 5 SNPs, only one pair of SNP information had an interaction of rs4855529:rs36330341 with the gene produced by INSR:FTO and a p-value of 0.0545. The p-value obtained is 0.1 and is positive, so the epistasis category is high.

G. Evaluation of Association Results

The current SVR model produces a fitness value of 15,09635. In other words, the current MAE value is 0.06624. This value is less satisfactory when compared to Tresnawati's [10] and Ramadhani's [15] research. Therefore, it is necessary to change the parameters in making the model. In this case, the parameter that is changed is the SVR kernel.

Initially, the kernel used was radial. The kernel will be changed to linear. Changing the kernel to linear has the advantage of finding an optimal weight that is useful for finding feature importance. After the kernel is replaced, the MAE obtained from the model is 0.02807. This result is much better than using a radial kernel. So, the model used in this research is the SVR model with a linear kernel with features selected from the previous SVR-GA. The model evaluation uses an MAE of 0.02807, where this result is better than Tresnawati's [10] study with an MAE value of 0.06222 and Ramadhani's research [15] with an MAE value of 0.06335.

Conclusion

This study succeeded in applying the Support Vector Regression and Genetic Algorithm methods for the association of SNPs to the insulin tolerance phenotype in T2D disease with an MAE value of 0.02807, where the MAE value is close to zero. The results obtained based on the SNP importance score were 14 SNPs. The association of SNP on insulin tolerance with the highest importance score is 0.25078 and SNPid rs30188721 chromosome 9, with a change in the A to G allele in the LIPC gene that instructs to make lipase enzymes that are passed into the blood. Epistatic highs for one SNP were rs30188721, rs32178367, and rs8277066 with LIPC, AKT2, and CAPN10 genes. Meanwhile, the epistasis for the two SNPs was rs36330341 with the FTO gene and rs48535529 with the INSR gene. The genes generated from the association can assist other researchers in finding the proper treatment for T2DM patients according to their genetic profile.

References

[1] RB Prasad and L Groop, "Precision medicine in type 2 diabetes"**.** *Journal of Internal Medicine,* 286(1):112-114. 2018 Dec 7. https://doi.org/10.1111/jiom.12859.

[2] J Ren, T He, Y Li, S Liu, Y Du, Y Jiang, C Wu, "Network-based regularization for high dimensional SNP data in the case–control study of Type 2 diabetes", *BMC genetics, 18(1), 44* 2017 Dec 18. https://doi.org/10.1186/s12863-017-0495-5

[3] C Sandor, NL Beer, C Webber. "Diverse type 2 diabetes genetic risk factors functionally converge in a phenotype-focused gene network", *PLoS computational biology*, *13*(10), p.e1005816. https://doi.org/10.1371/journal.pcbi.1005816

[4] S Leo, L Pireddu, G Zanetti. "SNP genotype calling with MapReduce", *MapReduce 12 - 3rd International Workshop on MapReduce and Its Applications:49–55.* 2017. doi: 10.1145/2287016.2287026.

[5] WQ Wei, JC Denny. "Extracting research-quality phenotypes from electronic health records to support precision medicine", *Genome medicine*, *7*(1), 41. 2015 Dec 7. https://doi.org/10.1186/s13073-015-0166-y

[6] MB Taylor, IM Ehrenreich. "Higher-order genetic interactions and their contribution to complex traits". *Trends Genet*. 2015;31(1):34-40. doi:10.1016/j.tig.2014.09.001

[7] BG Hall, "SNP-associations and phenotype predictions from hundreds of microbial genomes without genome alignments"**.** *PLoS One, 9(2), e90490.* 2014 Feb 28. https://doi.org/10.1371/journal.pone.0090490

[8] HJ Ban, JY Heo, KS Oh, KJ Park, "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine"**.** *PMID: 20416077; PubMed Central PMCID: PMC2875201*. 2010

https://doi.org/10.1186/1471-2156-11-26

[9] I Ilhan, G Tezel, "How to select tag snps in genetic association studies? The CLONTagger method with parameter optimization.", *a Journal of Integrative Biology.* 17(7):368–383. 2013. https://doi.org/10.1089/omi.2012.0100.

[10] LH Tresnawati, WA Kusuma, SH Wijaya, LS Hasibuan. "Asosiasi Single Nucleotide Polymorphism pada Diabetes Mellitus Tipe 2 Menggunakan Random Forest Regression", Jurnal Nasional Teknik Elektro dan Teknologi Informasi. 8(4):320-325. 2019 http://dx.doi.org/10.22146/jnteti.v8i4.531

[11] U Ilhan, G Tezel, C Özcan "*Tag SNP selection using similarity associations between SNPs*". *International Symposium on Innovations in Intelligent SysTems and Applications. Proceedings.* 2015 doi: 10.1109/INISTA.2015.7276793.

[12] MM Mukaka. "*Statistics corner: A guide to appropriate use of correlation coefficient in medical research*", *Malawi medical journal : the journal of Medical Association of Malawi, 24(3), 69–71.* 2012 https://doi.org/10.4236/jwarp.2015.77047

[13] B Coustet, P Dieudé, M Guedj, M Bouaziz, J Avouac, B Ruiz, J Sibilia, "C8orf13–BLK is a genetic risk locus for systemic sclerosis and has additive effects with BANK1: Results from a large french cohort and meta-analysis *Arthritis & Rheumatism*, *63*(7), 2091-2096. 2011. https://doi.org/ 10.1002/art.30379

[14] M Kayri, I Kayri and M. T Gencoglu, "The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data. *14th International Conference on Engineering of Modern Electric Systems (EMES)*, 2017, pp. 1-4, https://doi.org/ 10.1109/EMES.2017.7980368. 2017

[15] 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," 2020 International Conference on Computer Science and Its Application in Agriculture (ICOSICA), 2020, pp. 1-6, doi: 10.1109/ICOSICA49951.2020.9243208.