Asosiasi *Single Nucleotide Polymorphism* pada Diabetes Mellitus Tipe 2 Menggunakan *Random Forest Regression*

Lina Herlina Tresnawati[1], Wisnu Ananta Kusuma[2,5,*)], Sony Hartono Wijaya[3], Lailan Sahrina Hasibuan[4]

***Abstract*—Precision medicine can be developed by determining association between genomic data, represented by Single Nucleotide Polymorphism (SNP), and phenotype of diabetes mellitus type 2 (T2D). The number of SNP is actually very abundance. Thus, sorting and filtering the SNP is required before conducting association. The purpose of this paper was to associate SNP with T2D phenotypes. SNP ranking was conducted to choose significant SNPs by calculating importance score. Selected SNPs were associated with T2D phenotype using random forest regression. Moreover, the epistasis was also examined to show the interactions among SNPs affecting phenotype. This paper obtained 301 importance SNPs. Top ten SNPs have association with five T2D protein candidates. The evaluation results of the proposed models showed the Mean Absolute Error (MAE) of 0.062. This results indicate the success of random forest regression in conducting SNP and phenotype association and epistatic examination between two SNPs.**

***Intisari*—*Precision medicine* dapat dikembangkan dengan menentukan asosiasi antara data *genomic* yang direpresentasikan oleh *Single Nucleotide Polymorphism* (SNP) dan fenotipe dari penyakit diabetes mellitus tipe 2 (T2D). SNP adalah penanda yang berjumlah sangat banyak. Untuk itu, diperlukan proses pengurutan dan penapisan sebelum dilakukan asosiasi. Tujuan makalah ini adalah melakukan asosiasi SNP dengan fenotipe T2D. Pemeringkatan SNP dilakukan untuk memilih SNP yang signifikan berdasarkan *importance score*. SNP yang terpilih diasosiasikan dengan fenotipe T2D dan dilakukan pemeriksaan epistatis (interaksi antar SNP). Metode yang digunakan adalah *random forest regression*. Makalah ini menghasilkan 301 SNP yang signifikan. Sepuluh SNP terbaik memiliki asosiasi dengan lima buah kandidat protein T2D. Hasil evaluasi menunjukkan bahwa model asosiasi yang diusulkan memiliki nilai *Mean Absolute Error* (MAE) sebesar 0,062. Hasil evaluasi ini menunjukkan keberhasilan metode *random forest regression* dalam melakukan asosiasi antara SNP dan fenotipe T2D serta memeriksa epistatis antar dua buah SNP.**

***Kata Kunci*—Diabetes Mellitus Tipe 2, Epistatis, Pemetaan Asosiasi, *Random Forest Regression*, *Single Nucleotide Polymorphism*.**

[1,2,3,4] *Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Jl. Meranti Wing 20 Level V Kampus IPB Dramaga Bogor 16680 (tlp: 0251-8625584; fax:0251-8625584; e-mail:lina_herlina@apps.ipb.ac.id, ananta@apps.ipb.ac.id,sony@apps.ipb.ac.id,lailan.sahrina@apps.ipb.ac.id)*

[5] *Pusat Studi Biofarmaka Tropika Institut Pertanian Bogor, Jl. Taman Kencana No.3 Bogor16128 (tlp:0251-8373561; fax:0251-8347525; e-mail:ananta@apps.ipb.ac.id)*

**)Penulis korespondensi; e-mail:ananta@apps.ipb.ac.id*

I. PENDAHULUAN

Precision medicine adalah bidang ilmu pengobatan yang membuat strategi diagnosis yang tepat untuk pasien, dengan mempertimbangkan perbedaan gen individu, *microbiome*, lingkungan, sejarah medis keluarga, dan gaya hidup [1]. Secara umum, *precision medicine* terdiri atas data *genome, transcriptome, proteome, metabolome, phenome, interactome, microbiome, activity, environment, imaging, electronic health record* (EHR), dan sensor data [2]. Saat ini, *precision medicine* fokus mempelajari penyakit kompleks, seperti kanker, parkinsons, dan *diabetes mellitus* (DM). DM merupakan penyakit yang disebabkan oleh kurangnya sekresi insulin atau resistensi insulin [3].

Menurut *International Diabetes Federation* (IDF), pada tahun 2017 terdapat 425 juta orang penderita DM di dunia, yang diperkirakan pada 2045 akan menjadi 629 juta. Terdapat sekitar 90% dari semua penderita DM termasuk DM Tipe 2 (T2D). Penelitian penyakit dan obat pada manusia menjadi hal yang utama, tetapi pengaplikasian terhadap manusia terbentur permasalahan etika, legalitas, dan implikasi sosial, yang biasa disingkat dengan ELSI. Manusia dan hewan adalah contoh sistem kompleks, sehingga pengujian obat dan respons penyakit manusia dapat diteliti melalui perantara hewan. Salah satu hewan yang sering digunakan pada penelitian penyakit T2D adalah tikus, karena tikus memiliki kemiripan genetik dan fisiologis dengan manusia [4].

Keadaan biologis individu dapat disimpulkan dari beberapa tipe data *omic*, seperti *genomic* dan *metabolomics.* Profil *omic* berguna untuk memprediksi asosiasi gen terhadap penyakit [5]. *Genomic* direpresentasikan oleh data genotipe sebagai *marker* yang berkaitan dengan fenotipe penyakit T2D. Salah satu *marker* dari data genotipe yang memengaruhi ciri-ciri fenotipe adalah *Single Nucleotide Polymorphism* (SNP – dibaca 'snip') [6].

Penelitian sebelumnya sudah banyak membahas hal-hal terkait asosiasi SNP. Identifikasi SNP berasosiasi terhadap penyakit *Colorectal Cancer* (CRC). Sebanyak 22 SNP diidentifikasi menggunakan *Sequenom* MassARRAY, yang menghasilkan dua buah SNP signifikan terhadap penyakit CRC, yaitu rs1321311G>T pada gen CDKN1A dan rs10411210C>T pada gen RHPN2 [7]. Identifikasi asosiasi gen yang memiliki *high-dimensional data* menggunakan metode Markov Blanket (MB-TDT) dapat mengidentifikasi minimal SNP yang berasosiasi terhadap spesifik penyakit, jika dibandingkan dengan *exhaustive-search* [8]. Pembuatan konstruksi *Bayesian network* dan *genotype-phenotype inference* menggunakan *GWAS statistic*. Penelitian ini menghasilkan peluang asosiasi

trait sebuah penyakit terhadap *SNP-risk allel*, termasuk penyakit diabetes mellitus [9].

Pengembangan algoritme *Pre-conditioned Random Forest Regression* (PRFR) menggunakan ratusan SNP GWAS untuk memprediksi komplikasi radio terapi. Pada tahap ini, dilakukan pemeringkatan SNP menggunakan *random forest* untuk memperoleh *importance score* [10]. *Random forest* digunakan pada penentuan *importance score* untuk identifikasi SNP individu [11]. Seleksi SNP menggunakan *ranking* variabel dan *Sequential Forward Floating Selection* (SFFS) [12]. Penelitian ini berhasil melakukan *ranking* variabel terbaik menggunakan *random forest*, tetapi belum mempertimbangkan adanya epistatis (interaksi antar SNP). *Random forest* memiliki keunggulan dalam menangani jumlah variabel besar dan jumlah pengamatan yang sedikit.

Pada penelitian sebelumnya terkait asosiasi SNP, informasi yang dihasilkan biasanya antara satu buah SNP yang berhubungan dengan penyakit atau antara satu buah SNP dengan satu buah gen, serta belum mempertimbangkan adanya interaksi antar gen. Sementara itu, pada kondisi nyata, interaksi gen ini sangat penting, karena individu yang mempunyai penyakit T2D dapat dipengaruhi oleh satu gen saja ataupun dipengaruhi oleh beberapa gen. Interaksi antar gen ini dapat diketahui dengan adanya interaksi antar SNP, karena SNP digunakan sebagai referensi untuk menentukan rentang DNA yang diturunkan bersama sifat atau penyakit T2D. Jumlah SNP pada satu individu sangat banyak. Oleh karena itu, perlu dilakukan adanya pemeringkatan dan penapisan terlebih dahulu untuk memperoleh SNP signifikan terhadap penyakit T2D. Hasil penelitian masih menghasilkan asosiasi antara satu buah SNP terhadap satu buah gen. Sementara itu, pada kondisi nyata, bisa saja satu gen dipengaruhi oleh banyak SNP ataupun satu SNP dipengaruhi oleh banyak gen.

Pada makalah ini dilakukan asosiasi SNP terhadap penyakit T2D menggunakan *random forest regression* dengan mempertimbangkan adanya interaksi antar SNP. Asosiasi dilakukan dengan melakukan pemeringkatan SNP terlebih dahulu untuk memperoleh *importance score*. SNP yang terpilih diasosiasikan dengan fenotipe penyakit T2D dengan mempertimbangkan satu gen dapat dipengaruhi oleh beberapa SNP dan dilakukan pemeriksaan interaksi antar dua buah SNP yang berasosiasi (dikenal dengan istilah epistatis). Adapun kontribusi dari makalah ini yaitu sudah mempertimbangkan adanya epistatis. Analisis epistatis termasuk kompleks karena kombinasi pencariannya meningkat secara eksponensial [13]. Maka, dengan adanya pendekatan ini, dihasilkan peluang untuk mengatasi permasalan interaksi epistatis.

II. METODE

Penelitian yang dilakukan melalui enam tahapan, yaitu pengumpulan data SNP, praproses data, penyusunan peringkat SNP, pemetaan asosiasi, pemeriksaan interaksi SNP, analisis, dan evaluasi.

A. Pengumpulan Data SNP

Data SNP yang digunakan adalah data SNP gen tikus yang diambil dari situs *The Mouse Phenome Database* (MPD) dengan alamat *website* https://phenome.jax.org/. Dicari kueri SNP sebanyak 98 kandidat protein yang berhubungan dengan gen T2D. Data fenotipe yang digunakan adalah insulin *tolerance* tikus yang berhubungan dengan penyakit T2D [14]. *Database* yang digunakan adalah CGD-MDA1. *Database* ini merupakan *database* tikus *inbred* tahun 2014 yang terdiri atas 470.000 lokasi genom.

B. Praproses SNP

Praposes data dilakukan dengan *encode* data SNP. Pada sebuah populasi, SNP yang paling sering terjadi disebut alel mayor dan selainnya disebut alel minor. Masing-masing *haplotipe* (*h*) direpresentasikan dengan *string* biner untuk SNP *bi-allelic* $h = \{h1,2\ h2, \ldots, hm\}$, $hi \in \{0, 1\}$ [15]. Angka 0 merepresentasikan alel mayor, sedangkan angka 1 merepresentasikan alel minor. Pada *sequence* genotipe, informasi alel dibentuk oleh {A/A, A/T, A/C, A/G…G/C, G/T}. Jika genotipe *q* mempunyai *m* SNP, dapat direpresentasikan menjadi $q = \{q1, q2, \ldots, qm\}$, $qi \in \{0, 1, 2\}$. Pada makalah ini, pengkodean SNP menggunakan fungsi *recode* SNP *package scrime* pada R versi 3.6.0. Representasi nilai $gi \in \{0, 1, 2\}$ ditambahkan 1 menjadi $gi \in \{1, 2, 3\}$, dengan

1 : kedua alel merupakan mayor homozigot;
2 : kedua alel merupakan heterozigot; dan
3 : kedua alel merupakan minor homozigot.

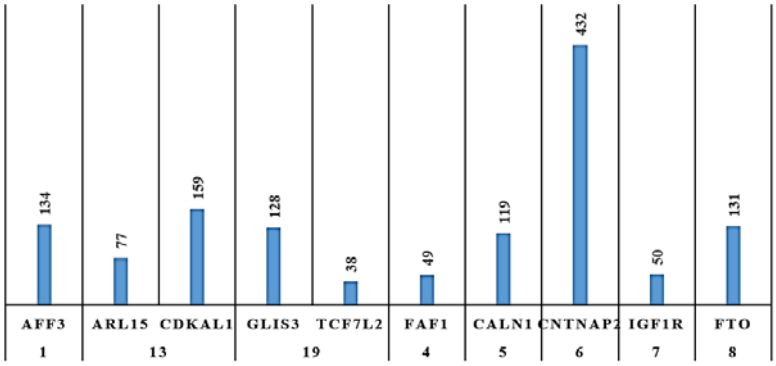
C. Penyusunan Peringkat SNP

Data penyususan peringkat SNP menggunakan algoritme *random forest* [16]. Algoritme ini dapat memperkirakan fitur SNP yang dianggap lebih penting. Hal ini mempermudah dalam pembuatan asosiasi. *Random forest* juga dapat memperkirakan perhitungan setiap atribut [17]. Penyusunan peringkat SNP dilakukan dengan menghitung penurunan nilai *Residual Sum of Square* (RSS) masing-masing SNP yang sudah dikodekan, menghitung, dan mengurutkan nilai *importance score*. Nilai *importance score* diambil dari urutan nilai terbesar sampai batas nilai nol. SNP yang memiliki nilai *importance score* lebih besar dibandingkan SNP yang lainnya dianggap mempunyai peluang asosiasi lebih besar terhadap fenotipe. Adapun tahapan untuk memperoleh *importance score* adalah sebagai berikut.

1. Untuk masing-masing SNP, *Mean Squared Error* (MSE) dihitung (yang dihitung pada masing-masing *tree* menggunakan *Out-of-Bag*/OOB data).
2. Satu SNP per *tree* pada OOB data dipermutasikan, sedangkan SNP yang tersisa dibiarkan tidak berubah. MSE yang dihasilkan dilambangkan dengan MSEp. MSEp ini biasanya lebih besar daripada MSE.
3. Langkah ini diulang untuk semua SNP selama proses pembangunan model.

D. Pemetaan Asosiasi

Pada tahap pemetaan asosiasi digunakan metode *random forest*. Adapun alasan dipilihnya metode *random forest* antara lain karena *random forest* dapat digunakan untuk memilih fitur terpenting [18]. Pada penelitian ini, SNP sebagai fiturnya. Selain itu, *random forest* dapat menjadi solusi untuk mengatasi permasalahan nonlinear. Analisis epistatis merupakan masalah kompleks dan nonlinear, serta komputasinya meningkat secara eksponensial [19]. Dengan demikian, penyelesaiannya dapat didekati menggunakan *random forest*.



Gbr. 1 Sepuluh besar jumlah kandidat protein pada setiap kromosom.

Random forest dikenalkan oleh Breiman pada tahun 2001. *Random forest* membangun *tree* menggunakan sampel *boostrap* data yang berbeda dan mengubah cara regresi membangun *tree*. Pada *tree* standar, setiap *node* dibagi menggunakan *split* terbaik di antara semua variabel, sedangkan pada *random forest* setiap *node* dibagi menggunakan yang terbaik di antara *subset* prediktor yang dipilih secara acak pada *node* tersebut. *Random forest* mempunyai dua buah parameter, yaitu jumlah variabel dalam *subset* acak di setiap *node* dan jumlah *tree* [20]. Adapun tahapan algoritme *random forest* adalah sebagai berikut.

1. Mendapatkan n_{tree} sampel *boostrap* dari data asli.
2. Membentuk *un-prunedregression tree* dengan modifikasi untuk masing-masing sampel *boostrap*, lalu mengambil prediktor secara acak dan memilih *split* terbaik di antara variabel-variabel tersebut.
3. Memprediksi data baru dengan menggabungkan jumlah *tree* (menggunakan rataan untuk regresi).

E. Pemeriksaan Interaksi SNP

Pemeriksaan interaksi SNP dilakukan dengan memeriksa susunan *tree* dari tahap penyusunan peringkat SNP. Jika SNP sering muncul pada susunan *tree random forest*, maka SNP tersebut diprediksi mempunyai nilai interaksi epistatis yang lebih besar. Parameter epistatis direpresentasikan dengan nilai *p-value*. Jika nilai *p-value* ≤ 0,1 dan positif, maka epistatis antar SNP termasuk *high*, jika nilai *p-value* ≤ 0,1 dan negatif, maka epistatis antar SNP termasuk *low*, dan epistatis SNP termasuk *undetermined* jika nilai *p-value*> 0,1.

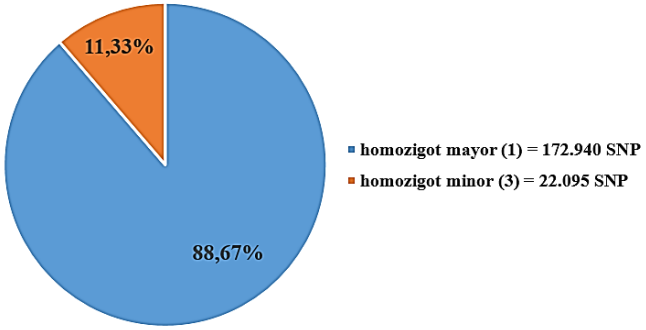
F. Analisis dan Evaluasi

Analisis dilakukan dengan membandingkan hasil asosiasi yang diperoleh dengan informasi asosiasi gen yang terdapat pada *Mouse Genome Informatics* (MGI) yang digabung dengan hasil penelusuran studi literatur. *Replication*, *Consortium*, dan *Epidemiology* mengamati alel yang signifikan berisiko T2D, bahkan pada SNP yang berisiko lemah [21]. Evaluasi dilakukan untuk melihat kinerja metode. Evaluasi model menggunakan *Mean Absolute Error* (MAE) yang merepresentasikan nilai *error* dari model asosiasi. Jika nilai MAE semakin mendekati nol, maka model tersebut semakin baik [22]. Persamaan MAE dapat dilihat pada (1).

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

TABEL I
RINGKASAN DATA DARI SITUS MPD

Komponen data	Keterangan
Genotipe	
Update	28 Maret 2019
Strain	10 *strain* tikus
Kromosom	1 – 19, X
Jumlah alel	Adenin (A) = 4.754 Timin (T) = 4.586 Guanin (G) = 5.669 Sitosin (C) = 5.521
Total SNP	2.053
Fenotipe	
Update	27 Februari 2019
Insulin tolerance	95 buah



Gbr. 2 Perbandingan hasil pengkodean SNP.

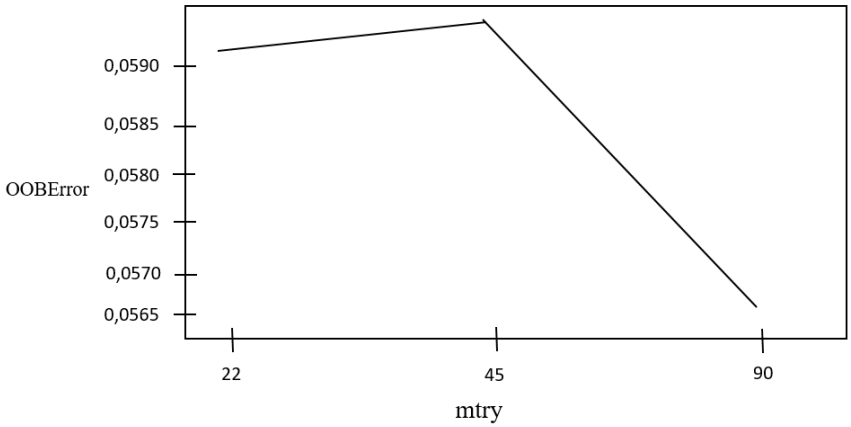
dengan

n = jumlah data
f_i = nilai hasil prediksi ke-i
y_i = nilai sebenarnya.

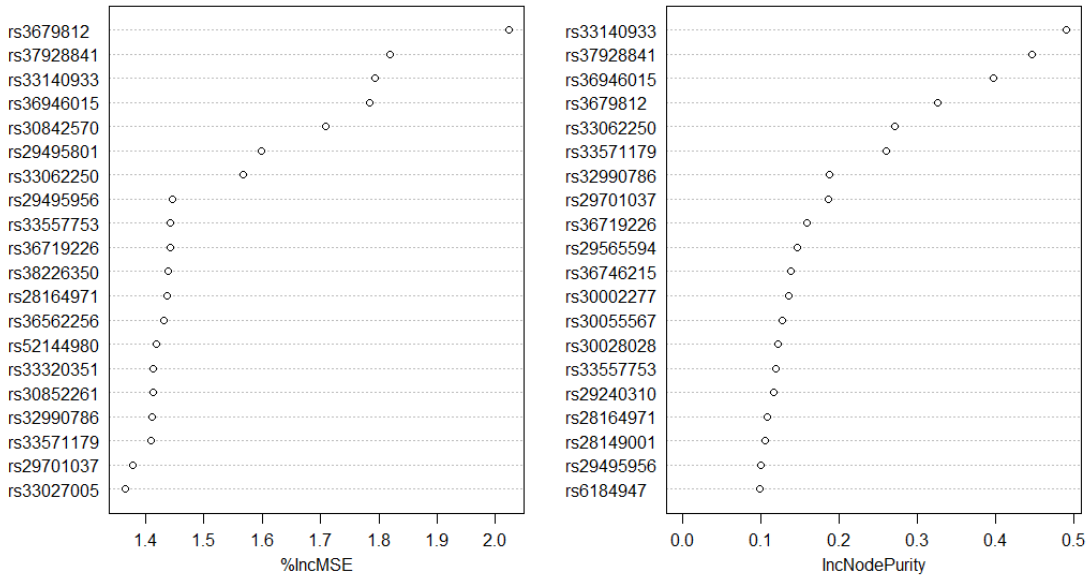
III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data SNP

MPD merupakan portal yang terdiri atas integrasi data *genomic* dan *phenomic* dengan memberikan akses ke data eksperimen primer, pengumpulan data yang sudah didokumentasikan, serta *tools* untuk analisis. Data pada situs ini dikontribusikan oleh peneliti dari seluruh dunia yang merepresentasikan karakteristik perilaku, kondisi morfologi dan fisiologi tikus, serta tikus yang terpapar obat atau



Gbr. 3 Nilai *OOBError* dan *mtry*.



Gbr. 4 Perbandingan dua puluh besar *SNP importance.*

perawatan lainnya. Ringkasan data yang diperoleh dari situs MPD dapat dilihat pada Tabel I.

Tabel I menunjukkan ringkasan data yang diperoleh dari situs MPD. Data terbanyak berupa alel Guanin (G). Adapun *strain* tikus yang digunakan adalah 129X1/SvJ, AKR/J, BALB/cByJ, C3H/HeJ, C57BL/6ByJ,C57BL/6J, DBA/1LacJ, NZB/BlNJ, PL/J, dan SWR/J. Semua *strain* tikus mempunyai sepuluh buah fenotipe *insulin tolerance*, kecuali *strain* AKR/J yang hanya mempunyai lima buah *insulin tolerance*. Sepuluh besar jumlah kandidat protein pada setiap kromosom dapat dilihat pada Gbr. 1. Gbr. 1 menunjukkan bahwa jumlah kandidat protein terbanyak adalah CNTNAP2, yaitu sebanyak 432 alel yang berada pada kromosom 6.

B. Praproses SNP

Hasil dari tahap praproses data SNP adalah sebagai berikut.

1. Jumlah *missing value* sebanyak 466 buah atau sekitar 22,7% dari total SNP. *Missing value* diisi menggunakan modus (nilai yang sering muncul) untuk setiap fitur SNP.
2. Pengkodean SNP menggunakan fungsi *recode*SNPs dalam *package scrime* pada R. Perbandingan hasil pengkodean SNP ditunjukkan pada Gbr. 2.
3. Data *insulin tolerance* digunakan semua dari *raw data*.

Gbr. 2 menunjukkan bahwa hasil pengkodean SNP didominasi oleh SNP homozigot mayor.

C. Penyusunan Peringkat SNP

Penyusunan peringkat SNP dilakukan menggunakan *random forest* berdasarkan persentase kenaikan MSE (*percent increase in mean squared error* atau %IncMSE saat nilai sebuah SNP dipermutasikan) dan penurunan (*decrease in node impurity* atau *increase node purity* atau %IncNodePurity akibat percabangan pada sebuah SNP). Gbr. 3 menunjukkan nilai *OOBError* yang diperoleh terhadap jumlah *mtry* setelah dilakukan *tuning* parameter. *Tuning* parameter menghasilkan nilai *mtry* optimal 90, dengan nilai *OOBError* 5,65%.

Jumlah *SNP importance* berdasarkan %IncMSE sebanyak 266 buah SNP, sedangkan jumlah *SNP importance*

TABEL II
TOP 20 HASIL ASOSIASI SNP TERHADAP *INSULIN TOLERANCE*

No	SNP ID	Kromosom	Posisi (bp38)	*Observed Allel*	Kandidat protein	*Importance score*
1	rs3679812	8	91.552.964	C/T	FTO	0,75204
2	rs33140933	5	130.388.947	A/T	CALN1	0,70049
3	rs33062250	5	130.413.203	C/G	CALN1	0,56114
4	rs33571179	5	124.264.260	A/G	MPHOSPH9	0,40175
5	rs36746215	8	91.535.609	C/T	FTO	0,37267
6	rs36895355	13	46.554.076	C/T	CAP2	0,34612
7	rs29695918	3	28.726.732	A/G	SLC2A2	0,31452
8	rs37928841	8	91.504.461	C/T	FTO	0,22288
9	rs47527350	6	115.378.014	C/T	PPARG	0,17688
10	rs6360050	13	113.863.702	C/T	ARL15	0,16847
11	rs45729233	5	130.676.538	C/T	CALN1	0,16513
12	rs29565594	5	124.324.722	C/T	MPHOSPH9	0,14807
13	rs29701037	3	28.728.279	C/G	SLC2A2	0,14135
14	rs29240310	13	29.333.294	A/T	CDKAL1	0,13419
15	rs29696739	3	28.708.248	C/G/T	SLC2A2	0,12505
16	rs37718141	5	130.744.782	C/T	CALN1	0,12262
17	rs33150625	8	91.492.318	A/C	FTO	0,12261
18	rs33650519	5	130.453.412	C/G	CALN1	0,11409
19	rs48207068	5	130.656.028	A/G	CALN1	0,11362
20	rs32996150	1	38.262.607	C/T	AFF3	0,10763

TABEL III
HASIL INTERAKSI DUA BUAH SNP

No	Pasangan SNP	*p-value*	Epistatis	Kandidat Protein
1	rs33643948:rs37967526	<2e-16 ***	*High*	CALN1 : FTO
2	rs33062250:rs33643948	1,47e-06	*High*	CALN1 : CALN1
3	rs33398625:rs37967526	6,41e-06	*High*	VIT : FTO
4	rs33643948:rs29494052	0,00103	*High*	CALN1 : CAP2
5	rs33062250:rs33398625	0,00108	*High*	CALN1 : VIT
6	rs29674897:rs29494052	0,01310	*High*	WFS1 : CAP2
7	rs29494052:rs37967526	0,01450	*High*	CAP2 : FTO
8	rs31000400:rs32978731	0,02161	*High*	PTPN22 : FTO
9	rs33062250:rs37967526	0,04510	*High*	CALN1 : FTO
10	rs33643948:rs31000400	0,05614	*High*	CALN1 : PTPN22
11	rs29674897:rs31000400	0,06933	*High*	WFS1 : PTPN22
12	rs33398625:rs29674897	0,08360	*High*	VIT : WFS1

berdasarkan %IncNodePurity jauh lebih banyak, yaitu 301 buah SNP. Karena jumlah *SNP importance* yang dihasilkan lebih banyak, SNP yang diambil untuk tahap selanjutnya adalah *SNP importance* berdasarkan %IncNodePurity. Perbandingan dua puluh besar *SNP importance* ditunjukkan pada Gbr. 4.

Pada Gbr. 4 terdapat satu buah *SNP importance* yang berada pada kedua hasil penyusunan peringkat SNP, yaitu rs37928841. Hal ini menunjukkan bahwa SNP tersebut mempunyai tingkat kepentingan dan peluang asosiasi yang lebih tinggi.

D. Pemetaan Asosiasi

Model asosiasi dilakukan menggunakan *random forest regression*. Setelah melakukan *tuning* parameter nilai *mtry*, *random forest* membangkitkan pohon keputusan dengan jumlah *tree* (*K*) = 50, dan *mtry* = 20. Pada model asosiasi ini, diperoleh nilai MAE sebesar 0,062. *SNP importance* yang diperoleh sebanyak 301 buah SNP. Dua puluh besar hasil asosiasi SNP terhadap fenotipe *insulin tolerance* disajikan pada Tabel II. Informasi yang diperoleh dari Tabel II adalah sebanyak dua puluh *SNP importance* yang berasosiasi terhadap *insulin tolerance* berhubungan dengan sembilan buah kandidat protein yang berhubungan dengan penyakit T2D.

E. Pemeriksaan Interaksi SNP

Pemeriksaan interaksi dua buah SNP dilakukan dengan menghitung kombinasi SNP yang terdapat pada *tree* pertama model asosiasi *random forest*. SNP yang berada pada *tree* yang sama berarti mempunyai nilai interaksi epistatis yang lebih besar. SNP hasil asosiasi menghasilkan 301 buah SNP, kemudian dikombinasikan dan dihitung nilai interaksinya menggunakan analisis varians (ANOVA) [23]. Dari 28 kombinasi SNP, terdapat sebanyak dua belas pasang SNP

epistatis *high*. Hasil interaksi pasangan SNP tersebut dapat dilihat pada Tabel III, yang jika pasangan SNP tersebut berinteraksi, maka akan menghasilkan asosiasi yang tinggi terhadap penyakit T2D.

F. Analisis dan Evaluasi

Evaluasi model *random forest regression* dilakukan menggunakan MAE. Nilai MAE yang diperoleh sebesar 0,062 dengan nilai *mtry* = 20 dan *K* = 50. Analisis menghasilkan informasi kandidat protein, kromosom, dan *observed allel* yang sesuai dengan MGI. Hasil penelusuran studi literatur dari tujuh buah kandidat protein menunjukkan bahwa sebanyak lima kandidat protein berhubungan dengan T2D (FTO, MPHOSPH9, SLC2A2, PPARG, ARL15) [24], [25]. Dari evaluasi model, *random forest regression* dapat digunakan pada proses penentuan SNP-SNP yang saling berinteraksi memengaruhi fenotipe T2D. Hal ini berimplikasi pada peningkatan efisiensi dan akurasi proses pemilihan dan penentuan asosiasi SNP dan fenotipe, khususnya yang terkait epistatis [26].

G. Perbandingan dengan Metode SVR

Seleksi SNP dilakukan menggunakan *ranking* variabel dan *Sequential Forward Floating Selection* (SFFS) yang membungkus *Support Vector Regression* (SVR). Hasil penelitian menghasilkan parameter optimal *kernel Radial Basis Function* (RBF) dengan nilai *cost* = 10. Hasil metode SVR ini kemudian diimplementasikan pada data SNP gen tikus MPD. Hasil yang diperoleh menunjukkan bahwa ternyata data SNP gen tikus dapat diimplementasikan menggunakan metode SVR, namun nilai MAE yang diperoleh sebesar 0,148. Nilai MAE ini masih lebih besar jika dibandingkan dengan nilai MAE penggunaan metode *random forest regression*, yaitu 0,062. Oleh karena itu, hal ini menunjukkan bahwa metode *random forest regression* lebih cocok digunakan untuk data SNP gen tikus MPD jika dibandingkan dengan metode SVR.

IV. KESIMPULAN

Makalah ini berhasil menerapkan metode *random forest regression* untuk asosiasi SNP terhadap fenotipe *insulin tolerance* pada penyakit T2D. Hasil yang diperoleh berupa pemeringkatan SNP dengan *importance score* sebanyak 301 SNP. Asosiasi SNP terhadap *insulin tolerance* terpenting terdapat pada kromosom 8, posisi 91.552.964 bp, perubahan alelnya dari C menjadi T, dengan kandidat protein FTO. Model ini mampu mendeteksi dua belas pasang SNP yang jika berinteraksi menghasilkan asosiasi yang tinggi terhadap penyakit T2D. Model yang dihasilkan dari pendekatan ini dapat digunakan untuk meningkatkan efisiensi dan akurasi proses penentuan asosiasi SNP dan fenotipe T2D, khususnya jika mempertimbangkan epistatis.

Adapun saran untuk bahan penelitian selanjutnya yaitu asosiasi SNP terhadap *insulin tolerance* dapat dilanjutkan menggunakan *multiple* fenotipe.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Kementrian Riset, Teknologi, dan Pendidikan yang telah menjadi sponsor studi magister melalui program Beasiswa PasTi (Beasiswa Pascasarjana Tenaga Kependidikan Berprestasi); Direktorat Sistem Informasi dan Transformasi Digital IPB; Pusat Studi Biofarmaka Tropika IPB; serta Kelompok Riset Bioinformatika Departemen Ilmu Komputer IPB.

REFERENSI

[1] X.D. Zhang, "Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships," *Journal of Pharmacogenomics Pharmacoproteomics*, Vol. 06, No. 02, hal. 1–2, 2015.

[2] B.E. Huang, W. Mulyasasmita, dan G. Rajagopal, "The Path from Big Data to Precision Medicine," *Expert Rev. Precis. Med. Drug Dev.*, Vol. 1, No. 2, hal. 129–143, 2016.

[3] Y. Yu, B. Wang, Z. Wang, F. Wang, dan L. Liu, "Wrapper Feature Selection Based Multiple Logistic Regression Model for Determinants Analysis of Residential Electricity Consumption," *2017 Asian Conf. on Energy, Power and Transport. Electrification (ACEPT)*, 2017, hal. 1-8.

[4] R.L. Perlman, "Mouse Models of Human Disease: An Evolutionary Perspective," *Evol. Med. Public Health*, Vol. 2016, No. 1 , hal. 170–176, 2016.

[5] K. Zarkogianni, M. Athanasiou, A.C. Thanopoulou, dan K.S. Nikita, "Comparison of Machine Learning Approaches Towards Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication," *IEEE J. Biomed. Heal. Informatics*, Vol. 22, No. c, hal. 1637-1647, 2017.

[6] A. Boutorh dan A. Guessoum, "Engineering Applications of Artificial Intelligence Complex Diseases SNP Selection and Classification by Hybrid Association Rule Mining and Artificial Neural Network — based Evolutionary Algorithms," *Eng. Appl. Artif. Intell.*, Vol. 51, hal. 58–70, 2016.

[7] B.W. Kang, H. Jeon, Y.S. Chae, S.J. Lee, J.Y. Park, J.E. Choi, J.S. Park, G.S. Choi, dan J.G. Kim, "Association between GWAS-Identified Genetic Variations and Disease Prognosis for Patients with Colorectal Cancer," *PLoS One*, Vol. 10, No. 3, hal. 1–9, 2015.

[8] H.J. Lee, J.W. Lee, S.H. Jin, H.J. Yoo, dan M. Park, "Detecting High-dimensional Genetic Associations using a Markov-Blanket in a Family-based Study," *2016 IEEE Int. Conf. on Bioinf. and Biomed. (BIBM)*, 2016, hal. 1767–1770.

[9] L. Zhang, Q. Pan, Y. Wang, X. Wu, dan X. Shi, "Bayesian Network Construction and Genotype-Phenotype Inference Using GWAS Statistics," *IEEE/ACM Trans. on Comp. Biol. and Bioinf.*, Vol. 16, No. 2, hal. 475-489, 2019.

[10] J.H. Oh, S. Kerns, H. Ostrer, S.N. Powell, B. Rosenstein, dan J.O. Deasy, "Computational Methods using Genome-wide Association Studies to Predict Radiotherapy Complications and to Identify Correlative Molecular Processes," *Nat. Publ. Gr.*, hal. 1–10, 2017.

[11] C. Yao, D.M. Spurlock, L.E. Armentano, C.D. Page Jr., M.J. Vandehaar, dan D.M. Bickhart, "Random Forests Approach for Identifying Additive and Epistatic Single Nucleotide Polymorphisms Associated with Residual Feed Intake in Dairy Cattle," *J. Dairy Sci.*, Vol. 96, No. 10, hal. 6716–6729, 2013.

[12] D. Setiawan, W.A. Kusuma, dan A.H. Wigena. "SNP Selection using Variable Ranking and Sequential Forward Floating Selection with Two Optimality Criteria," *J. Eng. Sci. Tech. Rev.*, Vol. 11, No. 5, hal. 76–85, 2018.

[13] L. Crawford, P. Zeng, S. Mukherjee, dan X. Zhou, "Detecting Epistasis with the Marginal Epistasis Test in Genetic Mapping Studies of Quantitative Traits," *PLoS Genetics*, Vol. 13, No. 7, hal. 1-37, 2017.

[14] C. Sandor, N. L. Beer, dan C. Webber, "Diverse Type 2 Diabetes Genetic Risk Factors Functionally Converge in a Phenotype-focused Gene Network," *PLoS Comput. Biol.*, Vol. 13, No. 10, hal. 1–23, 2017.

[15] U. Ilhan, G. Tezel, dan C. Özcan, "Tag SNP Selection Using Similarity Associations between SNPs," *Proc. 2015 Int. Symp. Innov. Intell. Syst. Appl. (INISTA)*, 2015, hal. 1-8.

[16] T.-T. Nguyen, J. Huang, Q. Wu, T. Nguyen, dan M. Li, "Genome-wide Association Data Classification and SNPs Selection Using Two-stage Quality-based Random Forests," *BMC Genomics*, Vol. 16, Suppl. 2, hal. 1-11, 2015.

[17] J.K. Jaiswal dan R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," *2017 World Congr. on Comp. and Comm. Tech. (WCCCT)*, 2017, hal. 65-68.

[18] K. Fawagreh, M.M. Gaber, dan E. Elyan, "Random Forests: From Early Developments to Recent Advancements," *Syst. Sci. Control Eng.*, Vol. 2, No. 1, hal. 602–609, 2014.

[19] X. Guo, Y. Meng, N. Yu, dan Y. Pan, "Cloud Computing for Detecting High-order Genome-wide Epistatic Interaction via Dynamic Clustering," *BMC Bioinformatics*, Vol. 15, No. 102, hal. 1–16, 2014.

[20] A. Liaw dan M. Wiener, "Classification and Regression by randomForest," *R News*, Vol. 2/3. hal. 18-22, 2002.

[21] A. Mahajan, M.J. Go, W. Zhang, J.E. Below, K.J. Gaulton, *et al*, "Genome-Wide Trans-ancestry Meta-analysis Provides Insight into the Genetic Architecture of Type 2 Diabetes Susceptibility," *Nat Genet.*, Vol. 46, No. 3, hal. 234-244, 2014.

[22] M. Kayri, I. Kayri, dan M.T. Gencoglu, "The Performance Comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using Photovoltaic and Atmospheric Data," *2017 14th Int. Conf. on Eng. of Modern Electric Systems (EMES)*, 2017, hal. 1-4.

[23] A. Wonkam, V.J.N. Bitoungui, A.A. Vorster, R. Ramesar, R.S. Cooper, B. Tayo, G. Lettre, dan J. Ngogang, "Association of Variants at BCL11A and HBS1L-MYB with Hemoglobin F and Hospitalization Rates among Sickle Cell Patients in Cameroon," *PLoS One*, Vol. 9, No. 3, hal. 1-9, 2014.

[24] S.A. Haddad, J.R. Palmer, K.L. Lunetta, dan M.C.Y. Ng, "A Novel TCF7L2 Type 2 Diabetes SNP Identified from Fine Mapping in African American Women," *PLoS One*, Vol. 12, No. 3, hal. 1–15, 2017.

[25] C.E. Arámbul-carrillo dan M.E. Ramos-márquez, "Association between Polymorphism in the AKT1 Gene and Type 2 Diabetes Mellitus in a Mexican Population," *Rev. Mex. Endocrinol. Metab. Nutr.*, Vol. 2, hal. 167–170, 2015.

[26] C.L. Schmalohr, J. Grossbach, M. Clément-Ziza, dan A. Beyer, "Detection of Epistatic Interactions with Random Forest Author Summary," *PLOS*, hal. 1–23, 2018.