PROCEEDINGS

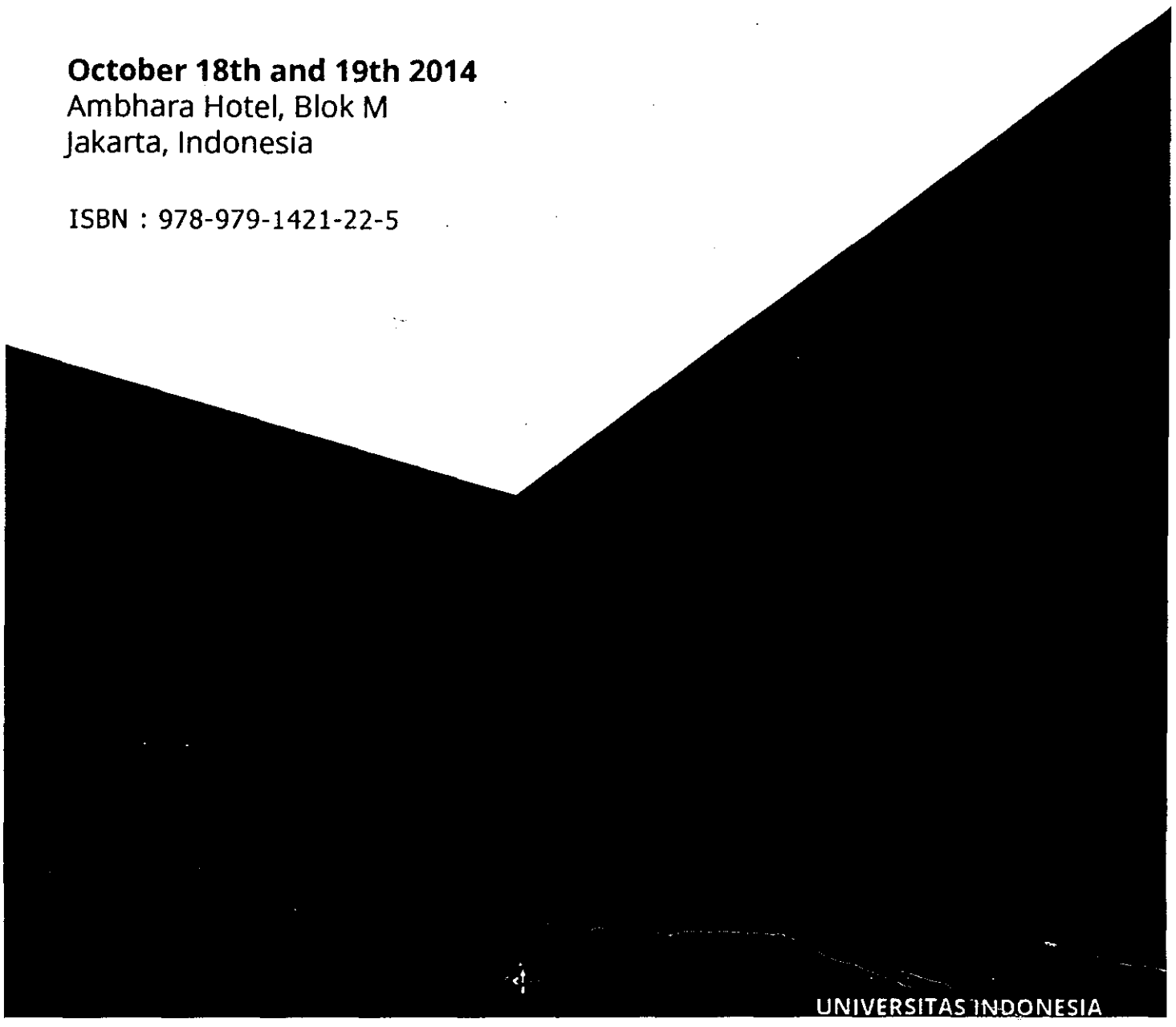
9

ICACSIS 2014

2014 International Conference on Advanced Computer Science and Information Systems

October 18th and 19th 2014
Ambhara Hotel, Blok M
Jakarta, Indonesia

ISBN : 978-979-1421-22-5



UNIVERSITAS INDONESIA

IEEE
INDONESIA SECTION

ICACSIS 2014

2014 International Conference on

Advanced Computer Science and Information Systems

(Proceedings)

ISBN : 978-979-1421-225

Welcome Message from General Chairs

On behalf of the Organizing Committee of this International Conference on Advanced Computer Science and Information Systems 2014 (ICACSIS 2014), we would like to extend our warm welcome to all of the presenter and participants, and in particular, we would like to express our sincere gratitude to our plenary and invited speakers.

This international conference is organized by the Faculty of Computer Science, Universitas Indonesia, and is intended to be the first step towards a top class conference on Computer Science and Information Systems. We believe that this international conference will give opportunities for sharing and exchanging original research ideas and opinions, gaining inspiration for future research, and broadening knowledge about various fields in advanced computer science and information systems, amongst members of Indonesian research communities, together with researchers from Germany, Singapore, Thailand, France, Algeria, Japan, Malaysia, Philippines, United Kingdom, Sweden, United States and other countries.

This conference focuses on the development of computer science and information systems. Along with 4 plenary and 2 invited speeches, the proceedings of this conference contains 71 papers which have been selected from a total of 132 papers from twelve different countries. These selected papers will be presented during the conference.

We also want to express our sincere appreciation to the members of the Program Committee for their critical review of the submitted papers, as well as the Organizing Committee for the time and energy they have devoted to editing the proceedings and arranging the logistics of holding this conference. We would also like to give appreciation to the authors who have submitted their excellent works to this conference. Last but not least, we would like to extend our gratitude to the Ministry of Education of the Republic of Indonesia, the Rector of Universitas Indonesia, Universitas Tarumanagara, Bogor Agricultural Institute, and the Dean of the Faculty of Computer Science for their continued support towards the ICACSIS 2014 conference.

Sincerely yours,
General Chairs

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

Welcome Message from The Dean of Faculty of Computer Science, Universitas Indonesia

On behalf of all the academic staff and students of the Faculty of Computer Science, Universitas Indonesia, I would like to extend our warmest welcome to all the participants to the Ambhara Hotel, Jakarta on the occasion of the 2014 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

Just like the previous five events in this series (ICACSIS 2009, 2010, 2011, 2012, and 2013), I am confident that ICASIS 2014 will play an important role in encouraging activities in research and development of computer science and information technology in Indonesia, and give an excellent opportunity to forge collaborations between research institutions both within the country and with international partners. The broad scope of this event, which includes both theoretical aspects of computer science and practical, applied experience of developing information systems, provides a unique meeting ground for researchers spanning the whole spectrum of our discipline. I hope that over the next two days, some fruitful collaborations can be established.

I also hope that the special attention devoted this year to the field of pervasive computing, including the very exciting area of wireless sensor networks, will ignite the development of applications in this area to address the various needs of Indonesia's development.

I would like to express my sincere gratitude to the distinguished invited speakers for their presence and contributions to the conference. I also thank all the program committee members for their efforts in ensuring a rigorous review process to select high quality papers.

Finally, I sincerely hope that all the participants will benefit from the technical contents of this conference, and wish you a very successful conference and an enjoyable stay in Jakarta.

Sincerely,
Mirna Adriani, Dra, Ph.D.
Dean of the Faculty of Computer Science
Universitas Indonesia

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

COMMITTEES

Honorary Chairs:

- A. Jain, IEEE Fellow, Michigan State University, US
- T. Fukuda, IEEE Fellow, Nagoya University, JP
- M. Adriani, Universitas Indonesia, ID

General Chairs:

- E. K. Budiarjo, Universitas Indonesia, ID
- D. I. Sensuse, Universitas Indonesia, ID
- Z. A. Hasibuan, Universitas Indonesia, ID

Program Chairs:

- H. B. Santoso, Universitas Indonesia, ID
- W. Jatmiko, Universitas Indonesia, ID
- A. Buono, Institut Pertanian Bogor, ID
- D. E. Herwindiati, Universitas Tarumanegara, ID

Section Chairs:

- K. Wastuwibowo, IEEE Indonesia Section, ID

Publication Chairs:

- A. Wibisono, Universitas Indonesia, ID

Program Committees:

- A. Azurat, Universitas Indonesia, ID
- A. Fanar, Lembaga Ilmu Pengetahuan Indonesia, ID
- A. Kistijantoro, Institut Teknologi Bandung, ID
- A. Purwarianti, Institut Teknologi Bandung, ID
- A. Nugroho, PTIK BPPT, ID
- A. Srivihok, Kasetsart University, TH
- A. Arifin Institut Teknologi Sepuluh Nopember, ID
- A. M. Arymurthy, Universitas Indonesia, ID
- A. N. Hidayanto, Universitas Indonesia, ID
- B. Wijaya, Universitas Indonesia, ID
- B. Yuwono, Universitas Indonesia, ID
- B. Hardian, Universitas Indonesia, ID
- B. Purwandari, Universitas Indonesia, ID
- B. A. Nazief, Universitas Indonesia, ID
- B. H. Widjaja, Universitas Indonesia, ID
- Denny, Universitas Indonesia, ID
- D. Jana, Computer Society of India, IN
- E. Gaura, Coventry University, UK
- E. Seo, Sungkyunkwan University, KR
- F. Gaol, IEEE Indonesia Section, ID

ISBN : 978-979-1421-225

- H. Manurung, Universitas Indonesia, ID
- H. Suhartanto, Universitas Indonesia, ID
- H. Sukoco, Institut Pertanian Bogor, ID
- H. Tolle, Universitas Brawijaya, ID
- I. Budi, Universitas Indonesia, ID
- I. Sitanggang, Institut Pertanian Bogor, ID
- I. Wasito, Universitas Indonesia, ID
- K. Sekiyama, Nagoya University, JP
- L. Stefanus, Universitas Indonesia, ID
- Marimin, Institut Pertanian Bogor, ID
- M. T. Suarez, De La Salle University, PH
- M. Fanany, Universitas Indonesia, ID
- M. Kyas, Freie Universitat Berlin, DE
- M. Nakajima, Nagoya University, JP
- M. Widyanto, Universitas Indonesia, ID
- M. Widjaja, PTIK BPPT, ID
- N. Maulidevi, Institut Teknologi Bandung, ID
- O. Sidek, Universiti Sains Malaysia, MY
- O. Lawanto, Utah State University, US
- P. Hitzler, Wright State University, US
- P. Mursanto, Universitas Indonesia, ID
- S. Bressan, National University of Singapore, SG
- S. Kuswadi, Institut Teknologi Sepuluh Nopember, ID
- S. Nomura, Nagaoka University of Technology, JP
- S. Yazid, Universitas Indonesia, ID
- T. Basaruddin, Universitas Indonesia, ID
- T. Hardjono, Massachusetts Institute of Technology, US
- T. Gunawan, International Islamic University Malaysia, MY
- T. A. Masoem, Universitas Indonesia, ID
- V. Allan, Utah State University, US
- W. Chutimaskul, King Mokut's University of Technology, TH
- W. Molnar, Public Research Center Henri Tudor, LU
- W. Nugroho, Universitas Indonesia, ID
- W. Prasetya, Universiteit Utrecht, NL
- W. Sediono, International Islamic University Malaysia, MY
- W. Susilo, University of Wollongong, AU
- W. Wibowo, Universitas Indonesia, ID
- X. Li, The University of Queensland, AU
- Y. Isal, Universitas Indonesia, ID
- Y. Sucahyo, Universitas Indonesia, ID

Local Organizing Committee:

- A. Y. Utomo, Universitas Indonesia, ID
- Aprinaldi, Universitas Indonesia, ID
- D. Eka, Universitas Indonesia, ID
- E. S. Utama, Universitas Indonesia, ID
- E. Cahyaningsih, Universitas Indonesia, ID
- H. A. Wisesa, Universitas Indonesia, ID
- H. R. Sanabila, Universitas Indonesia, ID
- K. M. Kurniawan, Universitas Indonesia, ID
- M. Mega, Universitas Indonesia, ID
- M. Ni'ma, Universitas Indonesia, ID

ISBN : 978-979-1421-225

- M. I. Suryani, Universitas Indonesia, ID
- M. Soleh, Universitas Indonesia, ID
- M. Pravitasari, Universitas Indonesia, ID
- N. Rahmah, Universitas Indonesia, ID
- Putut, Universitas Indonesia, ID
- R. Musfikar, Universitas Indonesia, ID
- R. Latifah, Universitas Indonesia, ID
- R. P. Rangkuti, Universitas Indonesia, ID
- V. Ayumi, Universitas Indonesia, ID

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

CONFERENCE INFORMATION

Dates	October 18th (Saturday) – October 19th (Sunday) 2014
Organizer	Faculty of Computer Science, Universitas Indonesia
	Department of Computer Science, Institut Pertanian Bogor
	Faculty of Information Technology, Universitas Tarumanegara
Venue	Ambhara Hotel
	Jalan Iskandarsyah Raya No. 1, Jakarta Selatan, DKI Jakarta, 12160, Indonesia
	Phone : +62-21-2700 888
	Fax : +62-21-2700 215
	Website : http://www.ambharahotel.com/
Official Language	English
Secretariat	Faculty of Computer Science, Universitas Indonesia
	Kampus UI Depok
	Depok, 16424
	Indonesia
	T: +62 21786 3419 ext. 3225
	F: +62 21 786 3415
	E: icacsis@cs.ui.ac.id
	W: http://www.cs.ui.ac.id
Conference Website	http://icacsis.cs.ui.ac.id

ISBN : 978-979-1421-225

PROGRAM SCHEDULE

Saturday, October 18th, 2014-CONFERENCE			
Time	**Event**	**Event Details**	**Rooms**
08.00-09.00	Registration		Dirgantara Room, 2nd Floor
09.00-09.30	Opening	Opening from the Dean of Faculty of Computer Science Universitas Indonesia/General Chair of ICACSIS 2014	
09.30-10.15	Plenary Speech I	Dr. Ir. Basuki Yusuf Iskandar, MA from Ministry of Communication and Information	
10.15-10.30	Coffee Break		
10.30-11.15	Plenary Speech II	Prof. Dame Wendy Hall from Southampton University, UK	
11.15.12.30	Lunch		
12.30-14.00	Parallel Session I : Four Parallel Sessions	**See Technical (Parallel Session I Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
14.00-15.30	Parallel Session II: Four Parallel Sessions	**See Technical (Parallel Session II Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
15.30-16.00	Coffee Break		
16.00-17.30	Parallel Session III : Four Parallel Sessions	**See Technical (Parallel Session III Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
17.30-19.00	Break		
19.00-22.00	Gala Dinner	Dinner, accompanied by music performance and traditional dances	Dirgantara Room, 2nd Floor

ISBN : 978-979-1421-225

ADVANCED PROGRAM ICACSIS 2014

Sunday, October 19th, 2014-CONFERENCE			
Time	**Event**	**Event Details**	**Rooms**
08.00-09.00	Registration		Dirgantara Room, 2nd Floor
09.00-10.00	Plenary Speech III	Drs. Harry Waluyo, M.Hum from Directorate General of Media, Design, Science & Technology Based Creative Economy	
10.00-10.15	Coffee Break		
10.15-11.30	Plenary Speech IV	Prof. Masatoshi Ishikawa from University of Tokyo, JP	
11.30-12.30	Lunch		
12.30-14.00	Parallel Session IV : Four Parallel Sessions	**See Technical (Parallel Session IV Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
14.00-15.30	Parallel Session V : Four Parallel Sessions	**See Technical (Parallel Session V Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
15.30-16.00	Coffee Break		
16.00-16.30	Closing Ceremony	Awards Announcement and Photo Session	Dirgantara Room, 2nd Floor

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

Table of Contents

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

Information Retrieval

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

Application of Decision Tree Classifier for Single Nucleotide Polymorphism Discovery from Next-Generation Sequencing Data

Muhammad Abrar Istiadi[1], Wisnu Ananta Kusuma[1], I Made Tasma[2]
[1]*Department of Computer Science. Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor, Indonesia*
[2]*Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD), Bogor, Indonesia*
E-mail: abrari@apps.ipb.ac.id, ananta@ipb.ac.id, imade.tasma@gmail.com

***Abstract*—Single Nucleotide Polymorphism (SNP) is the most abundant form of genetic variation and proven to be advantageous in diverse genetic-related studies. However, accurate determination of true SNPs from next-generation sequencing (NGS) data is a challenging task due to high error rates of NGS. To overcome this problem, we applied a machine learning method using C4.5 decision tree algorithm to discover SNPs from whole-genome NGS data. In addition, we conducted random undersampling to deal with the imbalanced data. The result shows that the proposed method is able to identify most of the true SNPs with more than 90% recall, but still suffers from a high rate of false-positives.**

***Keywords*- C4.5; decision tree; next-generation sequencing; single nucleotide polymorphism.**

I. INTRODUCTION

Single Nucleotide Polymorphism (SNP) is the simplest form of genetic variation among individuals [1]. It is defined as the mutation of single nucleotide base at specific points of individual's DNA sequence (as illustrated in Fig. 1). Despite its simplicity, SNP covers a large portion of variation and most of trait differences in a species, and is the most abundant form of genetic variation [2].

In the field of human genetic studies, analysis of SNPs and their effects are increasingly essential, for example in genetic analysis of disease, association genetics, pharmacogenomics, personalized medicine, and haplotype mapping [2]-[4]. SNP also plays important role in plant genetics, especially plant breeding, in which SNPs are used as molecular markers to facilitate more efficient and effective breeding scheme known as MAS (*marker assisted selection*) [5].

The implementation of SNP for genetic studies requires a fair amount of DNA sequence data to be analyzed. Correspondingly, recent advancement in sequencing technology has introduced methods to sequence individual's DNA in a high-throughput manner, known as next-generation sequencing (NGS) technology. NGS allows the DNA sequencing process to be faster, more cost-efficient, and produce significantly greater amount of data compared to conventional or traditional sequencing method [6]. In contrast, with massive amount of data that it produces, NGS has disadvantages that the data produced have relatively low quality and suffer from high error rates. These errors arise from multiple factors, namely base-calling errors introduced by the sequencing machine, and errors due to misalignment of sequence data [7], [8]. This limitation may impact subsequent downstream analysis, including SNP discovery, in which a false variant (a false SNP) may be called as true variant and vice versa.

```
                    SNP
Reference     ACCGTACACTAC

Sequence 1     CCTTAC
Sequence 2       GTAGACT
Sequence 3       GTACAC
Sequence 4         TAGACTCA
Sequence 5         TAGACTCAC
```

Fig. 1. Illustration of a multiple sequence alignment result [3]. The highlighted columns are positions where the nucleotide bases are polymorphic or have variations.

To address this problem, several methods based on machine learning have been proposed to distinguish between true and false variants. In [2], decision tree

ISBN : 978-979-1421-225

was constructed using C4.5 algorithm to classify true and false variants from non-NGS soybean sequence-tagged sites (STS) data. The decision tree method yielded improvement over SNP discovery without the incorporation of machine learning. Another approach using support vector machine (SVM) is described in [7] which employed human exome data sequenced on NGS platform. The SVM approach is shown to be effective combined with a number of features to determine true variants from an alignment data. In this work, we use C4.5 decision tree algorithm similar to [2] to discover SNP. We extend the decision tree using more features and apply the method on NGS whole-genome data, in which we differ from previous researches. We also try to deal with imbalanced dataset problem that arise from our findings.

II. METHODS

Preprocessing of the Sequence Data

To discover SNP from NGS sequence data, a reference sequence is required as a basis for aligning the short DNA reads produced by the sequencing platform [9], [10]. In this study, we use soybean (*Glycine max*) whole genome as the reference. The soybean genome consisting of 20 chromosomes (labeled Gm01 to Gm20) was assembled from cultivated soybean cultivar Williams 82 [11].

Raw sequence short reads were obtained from [12] which sequenced the whole genome from 14 accessions of cultivated soybean. The reads were paired-end and have identical length (75 base pairs). We conducted a quality control procedure to clean the data from ambiguous and low-quality bases [10] using PRINSEQ program, and then aligned all of the reads to the reference sequence using SOAP2 program in ungapped paired-end alignment mode. We chose SOAP2 in order to get similar alignment result with [12]. The alignment results in BAM format were then analyzed to extract the SNP information.

Feature Extraction

For each variant positions in the alignment result (the position where the base of reference is not equal with the bases of aligned reads), we extracted features combined from [2], [7], [9]. The complete list of features is listed in Table I and briefly described in the Appendix section. Note that some features have more than one value (for example major and minor allele), and one of the feature is nominal type (*variation type* feature), whereas the rest are numeric.

We assigned class label (*true* or *false*) to each variant by following strategy. A list of known soybean SNPs with their position and chromosome label from [12] was used as a "gold standard". If the current variant candidate (along with its position and chromosome label) is present in the list, then it is assigned to *true* class. Otherwise, it is considered as a false variant and assigned to *false* class. In this study, we did not cover variation in the type of small insertion and deletion (indels) as some of the features are not applicable for such variation type.

Classification

The generated training data from feature extraction step was classified using C4.5 algorithm [13]. We ran the experiments on Weka environment with **-M** parameter (minimum number of instances per leaf) set to 200 and tree pruning enabled to avoid overly complex model. The model is tested using 5-fold cross-validation and a number of test set taken from a subset of the total data. Performance of the model is measured using accuracy, true positives (TP) rate, and false positives (FP) rate [14].

For computational efficiency reason, we chose not to use all of the total data for training. Instead, the training data were taken from the longest chromosome of soybean (chromosome number 18, labeled Gm18), and the other soybean chromosome (Gm01, the second longest chromosome) were used to test the model against new instances. We also compared our model with previous method that also dealt with the same problem.

TABLE I
COMPUTED FEATURES FOR EACH VARIANT SITE

Feature	References
Variation type (transition-transversion)	[2]
Maximum quality of major and minor allele	[2]
Mean quality of major and minor allele	[2]
Frequency of major and minor allele	[2], [9]
Relative distance	[2], [9]
Mean base quality	[2], [7], [9]
Alignment depth	[2], [7], [9]
Alignment quality	[2], [7]
Distance to nearest variant	[9]
Error probability	[7]
Dinucleotide repeat	[7]
Mismatch area	[7]
Total mismatch count	[7]
Nucleotide diversity	[7]
Homopolymer length	[7]
Allele balance	[7]
Mean of nearby base quality	[7]

ISBN : 978-979-1421-225

TABLE II
SUMMARY OF CLASSIFICATION RESULTS

Dataset	Test Data	Accuracy (%)	TP Rate (%)	FP Rate (%)	Precision (%)
Gm18 imbalanced	5-fold CV	93.3	56.7	3.1	64.4
	Gm01	95.5	56.9	2.1	63.3
Gm18 random-undersampled	5-fold CV	89.5	92.8	13.8	87.1
	Gm01	89.5	92.2	10.7	35.0

III. RESULTS

Classification on Imbalanced Data

Across the whole genome, we found 39,454,648 SNP candidates, in which 2,823,603 candidates were assigned to *true* and the rest were assigned to *false*. This number led to the problem of class-imbalanced learning [14], since the positive or *true* class instances were just about 8% of the total instances. The percentage of *true* class varied between chromosomes, ranged from 5% to 10%. In our training data (chromosome Gm18) the *true* class was about 9% of the total 2,610,445 candidates on Gm18.

Using this imbalanced data, the classifier generated a rather complex tree with size of 405. The overall accuracy obtained by 5-fold cross-validation was about 93%. However, we observed that this accuracy was biased towards the *false* class (the majority class), since the TP rate of *true* class was low (56.7%) as well as the relatively low precision (64.4%). Additionally, we got a low number of false positives with FP rate of 3.1%. The test result of this model against a new dataset (chromosome Gm01) gave similar results as presented in Table II.

Random Undersampling to Rebalance the Dataset

We conducted a simple random undersampling [14] to reduce *false* class instances with the purpose of gaining a 1:1 class balance, thus retained only about 18% of Gm18 instances. With this reduced instances, the generated tree had less complexity with size of 187. The overall accuracy (89.5%) was smaller than the model generated using imbalanced data, but we achieved a significantly better TP rate (92.8%). The FP rate was somewhat worse compared to previous model (13.8%).

For precision, we gained a relatively high value by cross-validation testing. In contrast, upon testing the model with the imbalanced Gm01 dataset, we got poor precision (35%), although the other metrics (accuracy, TP rate, FP rate) were similar with the result of cross-validation (Table II).

Comparison with Previous Method

We tested our soybean data in the SNPSVM package [7] to compare our result with the result generated using SVM method. We used the same soybean chromosome for SVM training and testing (Gm18 and Gm01, respectively) without any modification or undersampling. The result from SVM was presented and compared in Fig. 2.

From Fig. 2, we could observe that our model using imbalanced dataset gave similar or slightly better result than SVM. Particularly for TP rate, our model performed significantly better (56.9% for C4.5 versus 37.8% for SVM).

We also compared the training time (not shown in the graph), and found that the C4.5 training took around 1.5 hours to complete, whereas the SVM needed considerably longer time to build the model (about 63.5 hours) on a standard workstation with 8-cores processor.

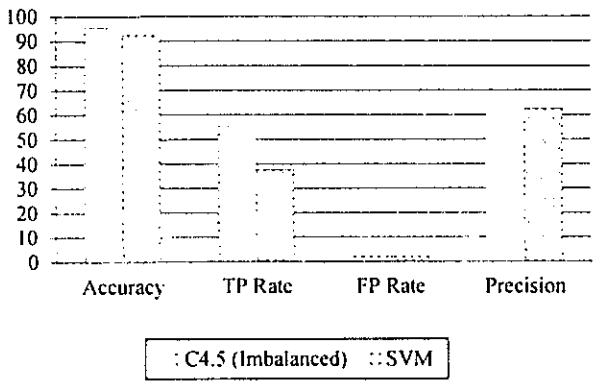


Fig. 2. Comparison of our results (using imbalanced data) and SVM result. All results were tested using Gm01 dataset.

IV. DISCUSSIONS

The problem of imbalanced class learning often arises in the analysis of molecular biology data [15]. In our study, any variation found was treated as SNP candidate (a data instance), even if the variation was just a single base of the entire reads aligned at the variant site. This candidate was often the result of misalignment or false reading from the sequencing machine, suggesting a false variant [8]. Accordingly,

ISBN : 978-979-1421-225

we predicted that this type of variant would naturally outnumber the true variants by some degrees, resulting in an imbalance between true and false variants.

Our experiment results using imbalanced data showed that the classifier was biased and only good at classifying the negative majority class (*false*), while missing almost half of the minority class (*true*). This behavior was expected (as the consequence of class-imbalanced learning [14]) but undesired, since the important true SNPs that might be useful for subsequent analysis would be unidentified. The low FP rate is the natural result of the biased classifier that is only good at negative class. Rebalancing the dataset gave a better classifier that significantly increased the rate of true positive calls, which would be able to identify most of the true SNPs. The performance comparison of the two approaches was further illustrated in Fig. 3, which clearly showed that classifier built on balanced data could outperform the biased classifier tested with cross-validation schema.

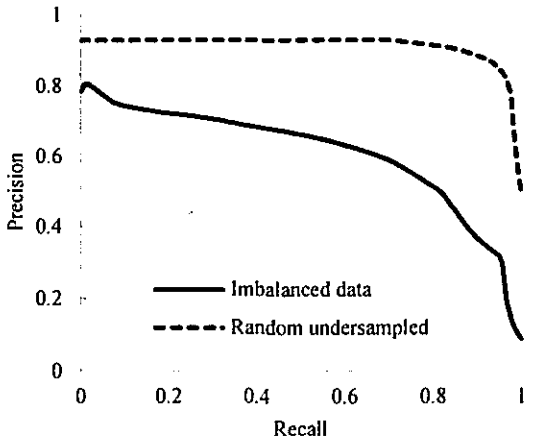


Fig. 3. Precision-recall curve for classification on imbalanced and random undersampled dataset (using cross-validation testing).

The comparison with SVM method showed that our C4.5 models could perform similar or better than SVM and with a significantly better training time. We also confirmed that SVM method suffered from imbalanced-class learning as well, denoted by the high accuracy but low TP rate (Fig. 2). The long training time of SVM showed one of the computational problem of NGS data analysis where large dataset have to be processed, hence a fast algorithm is essential.

Overall, the performance metrics of the models when tested against new instances were similar to the cross-validation testing, suggesting that the models were not overfitted. However, the precision of random-undersampled model was considerably low. This condition occurred because although the training data had been balanced, the class distribution of the test set (Gm01) was still imbalanced (6% positives). The imbalance introduced many false-positives (10.7%) that affected the precision of the model. Unfortunately, this rather high false-positives rate is also undesirable for analysis of SNPs that requires a high level of specificity [7].

V. Conclusion

From the results, we can conclude that our model trained from random-undersampled data performed arguably well on identifying the true variants (high recall), but still suffered from many false variants (low precision). Thus, more works are needed to deal with the nature of imbalanced dataset of SNP discovery, with the aim to achieve high recall as well as high precision with new instances.

Appendix

Brief description of computed features:

- Variation type: transition (A/G or T/C), otherwise transversion.
- Major and minor allele: major allele is the most common bases, and minor allele is the second most common. Here we calculate their maximum quality, mean quality, and bases frequency (divided by alignment depth)
- Relative distance: relative position of the variant site to both ends of the read and divided by read length
- Mean base quality: mean quality of all bases at the variant position.
- Alignment depth: number of reads aligned at variant position
- Alignment quality: quality given by the sequence alignment program
- Distance to nearest variant: distance of the variant to its neighboring variant (left and right flank size)
- Error probability: probability of the number of the reads containing variant base was sampled from binomial distribution with given parameters
- Dinucleotide repeat: number of dinucleotide repeat around the variant position (left and right direction)
- Mismatch area: mean number of variant base per each reads aligned at variant position
- Total mismatch count: number of mismatch on reads with reference base and reads with variant base
- Nucleotide diversity: deviation of reference base frequencies from given whole-genome average
- Homopolymer length: length of repeating bases around the variant position (left and right direction)
- Allele balance: number of reads containing variant bases divided by alignment depth
- Mean of nearby base quality: mean quality of all bases around variant position

Acknowledgment

The authors wish to thank Habib Rijzani and Dani Satyawan from ICABIOGRAD for the helpful

ISBN : 978-979-1421-225

discussions about SNP discovery, and Indonesia Ministry of Agriculture for funding this research through the KKP3N 2014 program.

REFERENCES

[1] B. S. Shastry, "SNPs: Impact on Gene Function and Phenotype." in *Single Nucleotide Polymorphisms*, A. A. Komar, Ed. Totowa, NJ: Humana Press, 2009, pp. 3–22.

[2] L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, I.-Y. Choi, P. B. Cregan, and C. P. Van Tassell, "Application of machine learning in SNP discovery," *BMC Bioinformatics*, vol. 7, no. 4, Jan. 2006.

[3] V. Bafna, A. Deutsch, A. Heiberg, C. Kozanitis, L. Ohno-Machado, and G. Varghese, "Abstractions for genomics," *Commun. ACM*, vol. 56, no. 1, p. 83, Jan. 2013.

[4] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation.," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–11, Jan. 2001.

[5] J. Mammadov, R. Aggarwal, R. Buyyarapu, and S. Kumpatla, "SNP markers and their impact on plant breeding," *Int. J. Plant Genomics*, vol. 2012, Jan. 2012.

[6] M. L. Metzker, "Sequencing technologies - the next generation.," *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, Jan. 2010.

[7] B. D. O'Fallon, W. Wooderchak-Donahue, and D. K. Crockett, "A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data," *Bioinformatics*, vol. 29, no. 11, pp. 1361–6, Jun. 2013.

[8] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, "Genotype and SNP calling from next-generation sequencing data.," *Nat. Rev. Genet.*, vol. 12, no. 6, pp. 443–51, Jun. 2011.

[9] J. Van Oeveren and A. Janssen, "Mining SNPs from DNA Sequence Data: Computational Approaches to SNP Discovery and Analysis," in *Single Nucleotide Polymorphisms*, vol. 578, A. A. Komar, Ed. Totowa, NJ: Humana Press, 2009, pp. 73–91.

[10] A. Altmann, P. Weber, D. Bader, M. Preuss, E. B. Binder, and B. Müller-Myhsok, "A beginners guide to SNP calling from high-throughput DNA-sequencing data," *Hum. Genet.* vol. 131, no. 10, pp. 1541–54, Oct. 2012.

[11] J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, *et al.*, "Genome sequence of the palaeopolyploid soybean.," *Nature*, vol. 463, no. 7278, pp. 178–83, Jan. 2010.

[12] H.-M. Lam, X. Xu, X. Liu, W. Chen, G. Yang, F.-L. Wong, *et al.*, "Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection," *Nat. Genet.*, vol. 42, no. 12, pp. 1053–9, Dec. 2010.

[13] J. Quinlan, *C4. 5: Programs for Machine Learning*, 1st edition. San Mateo: Morgan Kaufmann, 1993.

[14] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[15] P. Yang, Z. Zhang, B. B. Zhou, and A. Y. Zomaya, "Sample Subset Optimization for Classifying Imbalanced Biological Data," in *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science Volume 6635*, vol. 6635, J. Z. Huang, L. Cao, and J. Srivastava, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 333–344.

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225