Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2024, 2024:47
https://doi.org/10.28919/cmbn/8468
ISSN: 2052-2541

A TWO-STEP FEATURE SELECTION APPROACH FOR IDENTIFYING SNPs ASSOCIATED WITH COLORECTAL CANCER

JASON SEBASTIAN SULISTYAWAN[1,*], KELVIN JULIAN[1,*], GREGORIUS NATANAEL ELWIREHARDJA[2,3], KUNCAHYO SETYO NUGROHO[2], BENS PARDAMEAN[1,2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia

[2]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[3]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Copyright © 2024 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Colorectal Cancer (CRC) continues to be a significant cause of cancer-related illness and deaths worldwide. Single Nucleotide Polymorphism (SNP) identification and analysis can serve as a potential biomarker for early detection and personalized treatment. This study contributes to this ongoing discourse by employing bioinformatics methods, focusing on feature selection for SNP analysis related to CRC. Utilizing metaheuristic algorithms, particularly the Genetic Algorithm (GA), we implement a two-step feature selection method using Spatially Uniform ReliefF (SURF) and GA to identify key SNPs correlated with CRC, utilizing a dataset obtained from a prior study. Our comprehensive experiment successfully identifies previously established genes associated with CRC, while also revealing novel SNPs that warrant further investigation for validation.

[*]Corresponding author

E-mail address: jason.sulistyawan@binus.ac.id

Received February 03, 2024

Keywords: colorectal cancer; genetic algorithm; single nucleotide polymorphisms; feature selection; spatially uniform relief.

2020 AMS Subject Classification: 68W50, 90C59.

1. INTRODUCTION

Colorectal cancer (CRC) has emerged as a substantial global health challenge, ranking as the fourth most diagnosed and fatal cancer worldwide, following lung cancer, prostate cancer, and ovarian cancer [1]. Recent trends indicate a concerning shift in the age distribution of CRC diagnosis, with a notable increase in cases among individuals under 50 years old [2]. This alarming global pattern has been documented even in countries such as Indonesia, where instances of CRC diagnoses in individuals as young as 17 years old have appeared [3-6].

In tandem with this demographic shift, the urgency to address CRC is underscored by the absence of a known cure, especially for advanced stages, often culminating in prolonged and fatal outcomes [7]. The search for a cure for cancer, including CRC, is still ongoing through various research methods and approaches within the field of bioinformatics, aiming to broaden and improve the chances of finding an effective solution for this fatal disease [7-8].

Traditional single nucleotide polymorphisms (SNP) identification with conventional statistics and clinical trials shows significant weaknesses, characterized by high costs, long time periods, and large resource allocation [16]. Meanwhile, bioinformatic approaches such as machine learning, deep learning, and metaheuristic algorithms offer more efficient and scalable alternatives. In recent years, various strategies and approaches in bioinformatics have been tried to reveal the complexity of colorectal cancer and other malignancies [9-10] such as machine learning, computer vision, polygenic risk scores, and comprehensive analysis of SNPs [11-15]. In cases like CRC where they identified significant SNPs and genes associated with cancer [18-20].

In the landscape of metaheuristic methods, the Genetic Algorithm (GA) emerges as a method of choice to do SNP analysis on. With its global search capabilities, GA is exceptionally adept at navigating intricate interactions among SNPs, especially where these interactions are complex and not fully understood. Its searching ability also allows it to easily discover global optima, instead of being trapped in a local optima, thanks to its population-based search. Moreover, GA's synergy with other methods, such as the Hybrid Taguchi-Genetic Algorithm (HTGA), Genetic Algorithm with Ant Colony Optimization (GACO), FGSA, GASVeM, and various others that other studies have done, not only showcases its versatility but also underscores its position as a cornerstone in advancing state-of-the-art SNP analysis methodologies [20-23].

We also observed and tested that in most research, incorporating an extra step for selecting features before feeding the data into the GA significantly improves the algorithm's accuracy. This enhancement not only boosts the effectiveness of the GA itself but also opens up opportunities for refining and improving the overall feature selection process. Although prior algorithms have proven their ability to detect relevant SNPs, there is still room for combining the original GA approach with more biologically oriented statistical methods. One such avenue involves integrating feature selection methods specifically designed for biomedical data, like ReliefF-based algorithms including ReliefF, SURF, TURF, MultiSURF, and others [24].

In this study, we employ the Spatially Uniform ReliefF (SURF) technique to effectively reduce the number of features. The objective was to streamline the dataset before integrating it into the Genetic Algorithm. SURF has efficiently filtered relevant gene data and detected gene-gene interactions [25]. Subsequently, we introduced a novel approach, termed SURF-GA, combining the strengths of SURF and GA. The research contributions encompass the following aspects:

- Identified three genes that were identified within previous studies to be related to CRC by utilizing the proposed SURF-GA
- Validation of the effectiveness of SURF-GA in pinpointing crucial SNPs associated with CRC.
- Discovery of 2 new SNPs, whose clinical significance remains unidentified.

2. Materials and Methods

This section discusses in more detail the utilized SNP dataset, the developed two-step feature selection process named SURF-GA, and the conducted Analysis at the end. Where the proposed

SURF-GA consists of SURF, a pre-defined feature selection algorithm for genetic data [25], and an implementation of a Genetic Algorithm to identify SNPs that have an importance to CRC within our selected dataset. The following sections will then be discussed more as shown in Figure 1 Research Workflow.

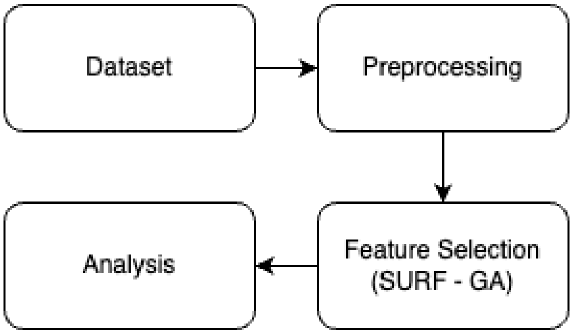

Figure 1. Research Workflow

2.1. DATASET

The dataset was collected from seven hospitals in the Makassar region of South Sulawesi, Indonesia. It was approved by the Hasanuddin University Ethical Committee and then managed by the Bioinformatics and Data Science Research Center (BDSRC) at Bina Nusantara University. The dataset includes medical records of participants, involving blood samples subjected to DNA extraction for genotyping using the Smokescreen Genotyping Array. The SNP data consists of 173 samples, 84 cases of CRC, and 89 cases of controls, along with the 446,395 SNP data collected. The following Table 1 displays a sample of the original SNP data displaying the case and control patient codes (C for Control, and CRC for Case) as the headers and the SNP locations as the rows.

Table 1. Sample SNP Data

index	**0_C118**	**0_CRC18**	**0_CRC77**
1:828166	0.707	0.463	0.181
1:830181	0.983	1.230	1.797
1:831489	0.960	1.185	1.845

2.2. PREPROCESSING

The initial dataset undergoes a preprocessing phase where we first restructure the data by transposing the data frame, creating a clearer format, showing clear correspondences between the columns and rows, and representing the SNPs, cases, and their respective values. Subsequently, each sample is assigned a target class based on pre-assigned class names, simplifying the format of the target column for enhanced clarity and simplicity. The following Table 2 displays the sample data that has gone through preprocessing.

Table 2. Preprocessed Sample Data

	1:82166	**1:830181**	**1:831489**	**CRC**
0_CRC26	0.146	1.856	1.946	1
0_C184	0.081	1.909	1.924	0
0_C148	0.012	1.985	1.989	0

2.3. SURF

Before feeding the preprocessed data into the Genetic Algorithm, a preliminary feature selection phase is undertaken to reduce the number of SNPs to 1000. This procedure is implemented to enhance the performance of the proposed GA. The feature selection process is conducted by using a pre-existing method, specifically tailored for identifying significant disease-related SNPs, known as SURF (Spatially Uniform ReliefF). SURF, an enhanced iteration of the ReliefF algorithm, is recognized for its efficacy in discerning important SNPs and SNP-SNP interactions [25].

Within this study, we utilized SURF by using the preprocessed dataset before, and feeding it to the provided SURF library from skrebate [24] as shown in Figure 2.

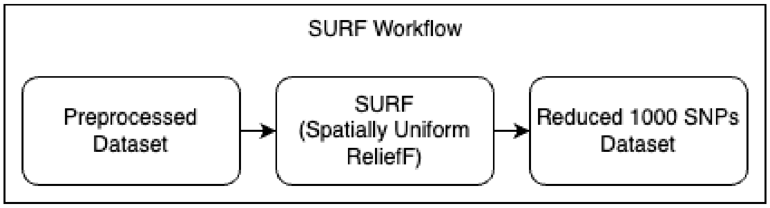

Figure 2. SURF Workflow

2.4. GENETIC ALGORITHM

Utilizing the 1000 selected SNPs, a Genetic algorithm is then implemented to identify crucial SNPs within this reduced dataset. The proposed Genetic Algorithm leverages the Area Under the Curve (AUC) score obtained from the Logistic Regression model, fitted on the selected SNPs, to establish the fitness values of individuals within the generation. Additionally, we introduce a novel crossover method, the Weighted Uniform Crossover, which will be expanded upon shortly. The flow of the proposed genetic algorithm is illustrated in Figure 3 under the Genetic Algorithm Workflow.

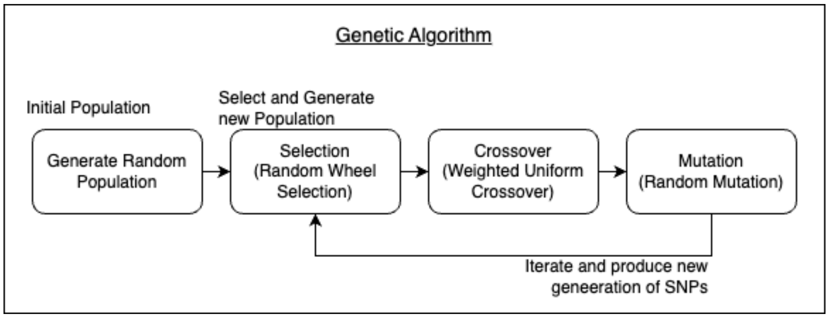
Genetic Algorithm
Initial Population
Select and Generate new Population
Generate Random Population
Selection (Random Wheel Selection)
Crossover (Weighted Uniform Crossover)
Mutation (Random Mutation)
Iterate and produce new geneeration of SNPs

Figure 3. Genetic Algorithm Workflow

The proposal of the Weighted Uniform Crossover method is motivated by a critical examination of previous studies, which predominantly rely on single, random, or simple uniform crossover techniques. Recognizing the potential for refinement in the crossover process, particularly to the uniform method, we advocate for the Weighted Uniform Crossover method. Diverging from the conventional Uniform crossover, this method introduces weighted considerations in the selection of individuals (specifically SNPs in this context) based on their fitness value using the following formula (1):

$$P(A) = F(A)F(A) + F(B) \qquad (1)$$

Where F(A) and F(B) represent the fitness score of individuals A and B respectively, and P(A) is the SNP crossover ratio from individuals A and B. This allows for offspring that inherit more SNP features from either parent depending on their fitness values. This implies that individuals with higher fitness values have a greater likelihood of passing on their SNP features to the next generation.

Regarding the mutation and selection strategies, we adhere to well-established implementations, as they are already deemed acceptable in previous studies [27-28]. The selected Random Wheel Selection method favors individuals with higher fitness values, elevating their chances of transitioning to the subsequent generation and potentially persisting throughout the iterative process. Meanwhile, for mutation, we adopt the Fixed Mutation Rate method, with each SNP feature in each individual being able to change into other randomly chosen SNP based on the set rate.

2.5. Analysis

Continuing the research process, after the Genetic Algorithm has found the important SNPs from the previously selected 1000 SNPs, we compare and validate the found SNPs' fitness to validate if the found SNPs were overfitted or could go through some other issues. At the start, we split the reduced dataset into two parts, train and test, for the analysis process itself, to validate the findings of the Genetic Algorithm proposed. Similar to the calculation of the fitness in the Genetic Algorithm, we also utilize a Logistic Regression model to which we then fit and validate the values of the top selected SNPs.

3. Experiment Results

3.1. Experiment Setup

The following tables display the parameters utilized in executing the experiment, where Table 3 displays the SURF parameters, and Table 4 displays the GA parameters.

Table 3. SURF Parameters

Parameter	**Value**
n_features_to_select	1000
verbose	1
n_jobs	-1

Table 4. GA Parameters

Parameter	**Value**
Generations	5000
Initial Population Count	300
Number of features / individual	4
Mutation rate	0.01

3.2. Experiment Results

The following Figure 4 displays the top 10 SNPs from the conducted SURF process before along with their feature importance scores and Table 5 shows the GA results and the top 5 individuals or sets of SNPs along with their test and fitness values derived from the conducted experiment. This experiment was executed on a Mac Mini equipped with an M1 processor, requiring approximately 10 hours. The overall outcomes yield meaningful findings regarding the identified significant SNPs that are potentially associated with CRC.

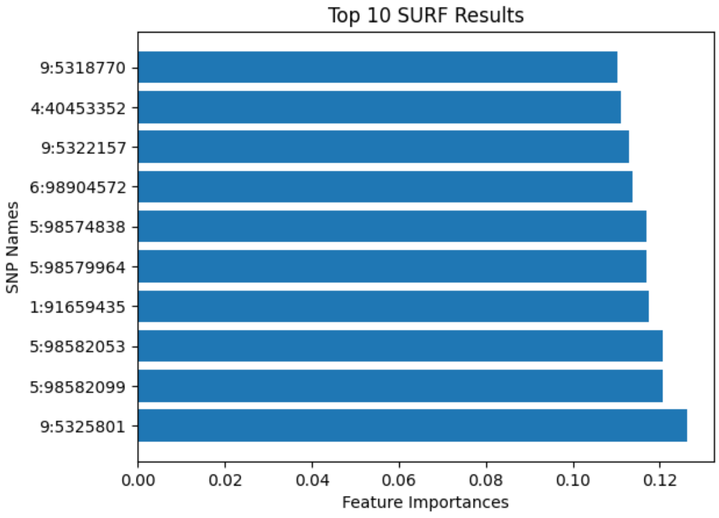

Figure 4. SURF Results

Table 5. GA Results

Test Score	Fitness Value	SNPs
84.2424	76.7399	[9:115495820, 5:4325504, 8:59894898, 18:5970704]
82.4242	75.3836	[7:54614383, 18:5974982, 1:203204169, 15:98818642]
80.0	75.3751	[7:54525340, 18:5965894, 12:718292884, 15:98818841]
78.7878	76.2709	[18:5974982, 10:36974772, 10:132431229, 15:98818642]
78.1818	78.1958	[18:5960165, 4:189633173, 19:39568654, 6: 98904572]

4. Discussion

4.1. SNP Findings

The SNPs discovered by the Genetic Algorithm are detailed in Table 6, presenting the top 20 sets of Single Nucleotide Polymorphisms (SNPs) generated by the GA. These SNPs have previously been associated with Colorectal Cancer (CRC) or other diseases, as documented in previous research and corroborated by the information available in the dbSNP from the National Library of Medicine site.

Table 6. Found SNPs with Clinical Importance

Location	SNP ID	Overlapping Gene
8:59894898, 8:59908664	rs79876400, rs115037695	TOX
18:5970704, 18:5965894, 18:5974982, 18:5960165, 18:5968696, 18:5968319	rs566559, rs692978, rs1106750, rs505559, rs950740323, rs1539807	L3MBTL4 and LOC121725015
7:54614383	rs2461636	VSTM2A
3:56238195	rs9880422	ERC2
8:47846757	rs7815490	LOC105375814

Based on the result, several identified SNPs overlapped with two distinct genes previously linked with Colorectal Cancer. Two SNPs were identified to overlap with the TOX gene, located at 8:59894898 and 8:59908664 aligning with previous research that implicates TOX as an enhancer for CRC development and a promoting factor for T cell exhaustion in human cancer [29-30] These findings suggest a potential regulatory role for these SNPs in influencing TOX gene function and contributing to CRC susceptibility. Furthermore, the identification of another SNP at 7:54614383 within the VSTM2A gene adds to existing evidence supporting its role as a suppressor of CRC through immune response modulation [31-32].

Aside from the previously identified SNPs within the TOX and VSTM2A genes, our comprehensive SNP analysis also identified three genes L3MBTL4, LOC121725015, and ERC2, each with distinct implications in human diseases. ERC2, for instance, has been previously linked to Maffucci's Syndrome in conjunction with other genes [33], with the identified SNP located at position 3:56238195. This discovery prompts further investigations in future research to validate the experimental findings. Two other genes, L3MBTL4 and LOC121725015, both located on chromosome 18, were associated with six SNPs collectively. Previous research has correlated L3MBTL4 with breast cancer and LOC121725015 with pancreatic cancer [34-35]. The substantial number of identified SNPs within these two genes suggests a need for in-depth reviews to validate the significance of these genes and their respective SNPs.

In addition to the aforementioned SNPs associated with known genes and diseases, our analysis identified several genes and SNPs that currently lack clinical evidence or association within the existing literature and dbSNP database [28]. Specifically, the gene LOC105375814, with the SNP at location 8:47846757, has yet to be clinically characterized. This gene presents an opportunity for further investigation to ascertain its potential relevance, particularly with colorectal cancer (CRC). Other notable SNPs discovered in our study, namely 8:47831255, 7:149010202, and 19:39567095, do not align with any known genes and lack clinical evidence in the current findings. Despite their unannotated status, these SNPs may represent novel genetic variations with implications for CRC or other diseases. Therefore, a more in-depth exploration of these SNPs is warranted in future research, as their discovery may contribute to expanding our understanding of genetic factors influencing disease susceptibility. The identification of these uncharacterized genes and SNPs underscores the potential for discoveries in the genetic landscape of CRC and highlights the importance of ongoing and comprehensive investigations in this field.

5. Conclusion

In this study, we utilized the SURF-GA methodology, which combines feature selection and a genetic algorithm, to identify the important single nucleotide polymorphisms (SNPs) associated with colorectal cancer (CRC). The investigation revealed notable SNPs, such as 8:59894898, 8:59908664, and 7:54614383, which align with past research linking them to the suppression or enhancement of colorectal cancer. Additionally, a cluster of SNPs on chromosome 18 emerged as significant, with previous findings suggesting associations with breast and pancreatic cancers. These results require further investigation to determine their correlation with CRC. Furthermore, novel SNPs, including 8:47831255, 7:149010202, and 19:39567095, were identified without precedent in existing research, potentially representing critical discoveries related to colorectal cancer. Nevertheless, these newly found SNPs require rigorous validation through further research and studies to substantiate their significance in CRC. Looking forward, the SURF-GA methodology, as a metaheuristic approach, holds promise, and future research endeavours should explore optimizations in selection, mutation, and crossover methods. Furthermore, the algorithm's applicability across various SNP datasets for different diseases should be systematically investigated to enhance its robustness and generalizability. This study lays the foundation for ongoing efforts in understanding the intricate genetic landscape of colorectal cancer and provides avenues for refining both methodologies and knowledge in other fields [36-38].

Author Contributions

All authors contributed equally to this study.

Conflict of Interests

The author(s) declare that there is no conflict of interests.

References

[1] Md.S. Hossain, H. Karuniawati, A.A. Jairoun, et al. Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies, Cancers. 14 (2022), 1732. https://doi.org/10.3390/cancers14071732.

[2] R.L. Siegel, K.D. Miller, A.G. Sauer, et al. Colorectal cancer statistics, 2020, CA: Cancer J. Clinicians. 70 (2020), 145-164. https://doi.org/10.3322/caac.21601.

[3] C.I. Pardamean, D. Sudigyo, A. Budiarto, et al. Changing colorectal cancer trends in Asians: Epidemiology and risk factors, Oncol. Rev. 17 (2023), 10576. https://doi.org/10.3389/or.2023.10576.

[4] D. Makmun, M. Simadibrata, M. Abdullah, et al. Colorectal cancer patients in a tertiary hospital in Indonesia: Prevalence of the younger population and associated factors, World J. Clin. Cases. 9 (2021), 9804-9814. https://doi.org/10.12998/wjcc.v9.i32.9804.

[5] M.R. Saraiva, I. Rosa, I. Claro, Early-onset colorectal cancer: A review of current knowledge, World J. Gastroenterol. 29 (2023), 1289-1303. https://doi.org/10.3748/wjg.v29.i8.1289.

[6] F. Iswandi, R.E. Asri, N.F. Lihawa, et al. Colorectal cancer in a 17-year-old boy: A case report, J. Med. Health Sci. J. 2 (2023), 55-62. https://doi.org/10.37905/jmhsj.v2i1.16816.

[7] A. Kumar, V. Gautam, A. Sandhu, et al. Current and emerging therapeutic approaches for colorectal cancer: A comprehensive review, World J. Gastrointest. Surg. 15 (2023), 495-519. https://doi.org/10.4240/wjgs.v15.i4.495.

[8] M.A. Horaira, M.A. Islam, M.K. Kibria, et al. Bioinformatics screening of colorectal-cancer causing molecular signatures through gene expression profiles to discover therapeutic targets and candidate agents, BMC Med. Genomics. 16 (2023), 64. https://doi.org/10.1186/s12920-023-01488-w.

[9] Y. Zhang, Y. Wang, B. Zhang, et al. Methods and biomarkers for early detection, prediction, and diagnosis of colorectal cancer, Biomed. Pharmacoth. 163 (2023), 114786. https://doi.org/10.1016/j.biopha.2023.114786.

[10] J.Y. Chao, H.C. Chang, J.K. Jiang, et al. Using bioinformatics approaches to investigate driver genes and identify BCL7A as a prognostic gene in colorectal cancer, Comput. Struct. Biotechnol. J. 19 (2021), 3922-3929. https://doi.org/10.1016/j.csbj.2021.06.044.

[11] T.W. Cenggoro, B. Mahesworo, A. Budiarto, et al. Features importance in classification models for colorectal cancer cases phenotype in Indonesia, Procedia Computer Sci. 157 (2019), 313-320. https://doi.org/10.1016/j.procs.2019.08.172.

[12] A.M.K. Izzaty, T.W. Cenggoro, G.N. Elwirehardja, et al. Multiclass classification of histology on colorectal cancer using deep learning, Commun. Math. Biol. Neurosci. 2022 (2022), 67. https://doi.org/10.28919/cmbn/7529.

[13] B. Mahesworo, A. Budiarto, B. Pardamean, Systematic evaluation of cross population polygenic risk score on colorectal cancer, Procedia Computer Sci. 179 (2021), 344-351. https://doi.org/10.1016/j.procs.2021.01.015.

[14] S. Amadeus, T.W. Cenggoro, A. Budiarto, et al. A design of polygenic risk model with deep learning for colorectal cancer in multiethnic Indonesians, Procedia Computer Sci. 179 (2021), 632-639. https://doi.org/10.1016/j.procs.2021.01.049.

[15] J.P. Trinugroho, A.A. Hidayat, M. Isnan, et al. Machine learning approach for single nucleotide polymorphism selection in genetic testing results, Procedia Computer Sci. 227 (2023), 46-54. https://doi.org/10.1016/j.procs.2023.10.501.

[16] L. Tong, B. Tayo, J. Yang, et al. Comparison of SNP-based and gene-based association studies in detecting rare variants using unrelated individuals, BMC Proc. 5 (2011), S41. https://doi.org/10.1186/1753-6561-5-s9-s41.

[17] J.S. Lin, L.A. Perdue, N.B. Henrikson, et al. Screening for colorectal cancer, JAMA. 325 (2021), 1978-1998. https://doi.org/10.1001/jama.2021.4417.

[18] A. Cakmak, H. Ayaz, S. Arıkan, et al. Predicting the predisposition to colorectal cancer based on SNP profiles of immune phenotypes using supervised learning models, Med. Biol. Eng. Comput. 61 (2022), 243-258. https://doi.org/10.1007/s11517-022-02707-9.

[19] B. Mahesworo, A. Budiarto, A.A. Hidayat, et al. Cancer risk score prediction based on a single-nucleotide polymorphism network, Healthc. Inform. Res. 28 (2022), 247-255. https://doi.org/10.4258/hir.2022.28.3.247.

[20] R.E. Caraka, M. Tahmid, R.M. Putra, et al. Analysis of plant pattern using water balance and cimogram based on oldeman climate type, IOP Conf. Ser.: Earth Environ. Sci. 195 (2018), 012001. https://doi.org/10.1088/1755-1315/195/1/012001.

[21] T.W. Cenggoro, F. Tanzil, A.H. Aslamiah, et al. Crowdsourcing annotation system of object counting dataset for deep learning algorithm, IOP Conf. Ser.: Earth Environ. Sci. 195 (2018), 012063. https://doi.org/10.1088/1755-1315/195/1/012063.

[22] I.D. Kurniawan, R.C.H. Soesilohadi, C. Rahmadi, et al. The difference on Arthropod communities' structure within show caves and wild caves in Gunungsewu Karst area, Indonesia, Ecol. Environ. Conserv. 24 (2018), 72-81.

[23] K. Muchtar, F. Rahman, T.W. Cenggoro, et al. An improved version of texture-based foreground segmentation: Block-based adaptive segmenter, Procedia Computer Sci. 135 (2018), 579-586. https://doi.org/10.1016/j.procs.2018.08.228.

[24] R.E. Caraka, S. Shohaimi, I.D. Kurniawan, et al. Ecological show cave and wild cave: Negative binomial Gllvm's arthropod community modelling, Procedia Computer Sci. 135 (2018), 377-384. https://doi.org/10.1016/j.procs.2018.08.188.

[25] C.S. Greene, N.M. Penrod, J. Kiralis, et al. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions, BioData Mining 2 (2009), 5. https://doi.org/10.1186/1756-0381-2-5.

[26] I. Yusuf, B. Pardamean, J.W. Baurley, et al. Genetic risk factors for colorectal cancer in multiethnic Indonesians, Sci. Rep. 11 (2021), 9988. https://doi.org/10.1038/s41598-021-88805-4.

[27] R.E. Caraka, M. Noh, R.C. Chen, et al. Connecting climate and communicable disease to penta helix using hierarchical likelihood structural equation modelling, Symmetry. 13 (2021), 657. https://doi.org/10.3390/sym13040657.

[28] N. Dominic, Daniel, T.W. Cenggoro, Arif Budiarto, Bens Pardamean, Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification, Commun. Math. Biol. Neurosci. 2021 (2021), 39. https://doi.org/10.28919/cmbn/5565.

[29] T. Chen, Q. Li, X. Zhang, et al. TOX expression decreases with progression of colorectal cancers and is associated with CD4 T-cell density and Fusobacterium nucleatum infection, Human Pathol. 79 (2018), 93-101. https://doi.org/10.1016/j.humpath.2018.05.008.

[30] K. Kim, S. Park, S.Y. Park, et al. Single-cell transcriptome analysis reveals TOX as a promoting factor for T cell exhaustion and a predictor for anti-PD-1 responses in human cancer, Genome Med. 12 (2020), 22. https://doi.org/10.1186/s13073-020-00722-9.

[31] Y. Dong, Y. Zhang, W. Kang, et al. VSTM2A suppresses colorectal cancer and antagonizes Wnt signaling receptor LRP6, Theranostics. 9 (2019), 6517-6531. https://doi.org/10.7150/thno.34989.

[32] Y. Dong, Y. Zhou, N. Wong, et al. Abstract 2332: VSTM2A modulation of immune response in colorectal cancer through abrogating PD-L1 and PD-1 interaction, Cancer Res. 83 (2023), 2332-2332. https://doi.org/10.1158/1538-7445.am2023-2332.

[33] P. Cheng, K. Chen, S. Zhang, et al. IDH1 R132C and ERC2 L309I mutations contribute to the development of Maffucci's syndrome, Front. Endocrinol. 12 (2021), 763349. https://doi.org/10.3389/fendo.2021.763349.

[34] L. Addou-Klouche, J. Adélaïde, P. Finetti, et al. Loss, mutation and deregulation of L3MBTL4 in breast cancers, Mol. Cancer, 9 (2010), 213. https://doi.org/10.1186/1476-4598-9-213.

[35] Y. Han, K.J. Jung, U. Kim, et al. Non-invasive biomarkers for early diagnosis of pancreatic cancer risk: metabolite genomewide association study based on the KCPS-II cohort, J. Transl. Med. 21 (2023), 878. https://doi.org/10.1186/s12967-023-04670-x.

[36] J. Baurley, A. Perbangsa, A. Subagyo, et al. A web application and database for agriculture genetic diversity and association studies, Int. J. Bio-Sci. Bio-Technol. 5 (2013), 33-42. https://doi.org/10.14257/ijbsbt.2013.5.6.04.

[37] M.F. Kacamarga, B. Pardamean, H. Wijaya, Lightweight virtualization in cloud computing for research, Commun. Computer Inform. Sci. (2015), 439-445. https://doi.org/10.1007/978-3-662-46742-8_40.

[38] J.W. Baurley, A. Budiarto, M.F. Kacamarga, et al. A web portal for rice crop improvements, Int. J. Web Portals 10 (2018), 15-31. https://doi.org/10.4018/ijwp.2018070102.