Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2023, 2023:51

https://doi.org/10.28919/cmbn/7962

ISSN: 2052-2541

SNP DISTRIBUTED REPRESENTATION USING ENTITY EMBEDDING

FRANCISCO CALVIN ARNEL FERANO[1], JONATHAN CHRISTIAN SETYONO[1], ARDIVO VIRSA SISWANTO[1], NICHOLAS DOMINIC[1,2], BENS PARDAMEAN[1,2]

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

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

Abstract. A single Nucleotide Polymorphism (SNP) array is the largest variation of genetic information to detect specific traits in organisms. SNP is located in a specific locus of DNA sequences. To the day this study was conducted, the representation of SNPs for machine learning models is still questionable. Based on the previous works, we proposed a comparative study of distributed representation methods against SNPs data. This study used 1,232 SNPs from the genomic data of 687 Indonesian rice samples collected from four distinct rice fields. The SNP data used was converted into an encoded format. Entity embedding (Embedder) and several comparative models, i.e., Node2Vec, Struc2Vec, and LINE, were chosen to predict the rice yield of the SNP data. The entity embedding using Embedder outperformed the comparative methods used in this study, namely Node2Vec, Struc2Vec, and LINE with the best R2 and MSE scores of 0.9368 and 0.2425 respectively.

Keywords: single nucleotide polymorphism; distributed representation; entity embedding; graph embedding; neural network.

2020 AMS Subject Classification: 92B20.

1. INTRODUCTION

Single Nucleotide Polymorphism (SNP) array is the largest variation of genetic information to

*Corresponding author

E-mail address: francisco.ferano@binus.ac.id

Received March 19, 2023

detect specific traits in organisms, such as colorectal cancer [1] and rice yield rate [2]–[5]. Biologically, SNP is located in a specific position (locus) of DNA sequences [6]–[8]. The polymorphism (which means “many forms”) or variation at each locus lies across the population. Therefore, SNP data cannot be treated the same as the DNA sequence, even though the SNP itself is a unique part of DNA. A high-performance data storing system, as well as an appropriate way of processing this data, are required due to the vast volume of SNP data [9]. The data used in this study was genomic data from 687 Indonesian rice samples taken from four different rice fields with a total of 1,232 SNPs. This resulted in a lot of SNP combinations for each rice sample. Hence, the sparseness of the data is tackled by using a representation technique to enhance the model performance to predict the rice yield from the rice SNP data.

To the day this study was conducted, the representation of SNPs for machine learning models is still questionable. In language modeling (LM), there are localist and distributed representations. The localist representation (LR), such as one-hot encoding for discrete random variables [10], suffers from its long sparse vector that requires n-bits as the representative of the corpus with *n*-words. On the other hand, the distributed [11] representations (DR) are denser as the embedding dimension can be freely set regardless of the n-words. DR is also context-dependent [11], contains continuous values instead of binary, and can generalize well [12] given the new unseen data. Consequently, this research used DR as the representation technique.

There are existing works regarding DR in LM such as entity embedding. According to Cheng Guo and Felix Berkhahn [13], entity embedding can create a generalized model when using sparse data. Entity embedding has three advantages: reducing memory usage, faster encoding than one-hot encoding, and revealing properties intrinsically by mapping the categorical variable [13]. This resulted in better machine learning performance when using neural networks to predict classification with a Mean Absolute Percentage Error (MAPE) from 0.101 to 0.093. Additionally, Feng Hou et al. [14] examined the problem of identifying links between embedding entities and proposed that FGS2EE attaches semantic information into entity embeddings to help model training converge faster by learning the contextual commonality of entities. The other study (Moreno et al., 2017) also examined the same problem. Davide Mezzogori and Francesco

Zammori utilized Deep Neural Networks (DNN) on Entity Embeddings for demand forecasting [15]. The proposed forecasting model achieved a MAPE within the range of 10 to 15%. Based on previous works, this research proposed a distributed representation of SNPs data using an entity embedding technique to predict the rice yield of the SNP data. Other research used entity embedding as can be seen in [16], [17].

Graph Embedding method can be used to create distributed representations such as DeepWalk, Large-scale Information Network Embedding (LINE), Node2Vec, Structural Deep Network Embedding (SDNE), and Struc2Vec. The first graph embedding method, DeepWalk proposed by Perozzi et al. [18], is an unsupervised feature learning to study latent representations in a graph network using graphs created from word sequences. Multi-label classification tasks for social network data are used for experiments such as BlogCatalog, Flickr, and YouTube. The proposed DeepWalk method achieved a 10% higher F1 score performance and outperformed other methods' performance using less training data Perozzi et al. [18]. The second one, Large-scale Information Network Embedding (LINE) was introduced by Tang et al. [19] to overcome problems in existing graph embedding methods which are graphs or networks that contain millions of nodes in the real world. According to Ma and Zhang [19], the proposed method of LINE is very efficient when embedding a network with billions of edges. This method improves the overall embedding effectiveness using Wikipedia data from 43.65% using DeepWalk to 66.10% using LINE with approximately 14 hours less training time than DeepWalk. Node2Vec, as the third method proposed by Grover and Leskovec [20], studied continuous feature representation in a graph and presented a more efficient way for multi-label classification. According to Grover and Leskovec [20], the Node2Vec method achieved an increase of up to 21.8% macro F1 score using Wikipedia from the previous graph embedding method such as DeepWalk and LINE. Then, Structural Deep Network Embedding (SDNE) is introduced by Wang et al. [21] with the method of semi-supervised deep model graph or network embedding. According to Wang et al. [21], SDNE achieved a better F1 score performance on multi-label classification tasks using Blogcatalog, Flickr, and YouTube data. The last method, Struc2Vec, proposed by Ribeiro et al. [22] has an advantage over other graph embedding methods which capture the structural identity of a graph or

network. Other than Tang et al. [19], there are a few other methods that used 2Vec such as [23]–[27].

In this study, entity embedding has been proposed to perform SNP embedding. Entity embedding was chosen because it has several advantages. When the input is sparse and the statistics are uncertain, entity embedding aids the neural network in generalizing more effectively. It also can reduce memory usage and perform faster encoding than one-hot encoding. This study employed 1,232 SNPs from the genomic data of 687 Indonesian rice samples collected from four distinct rice fields. This resulted in a lot of SNP combinations for each rice sample.

The rest of this paper is organized as follows. In Section 2, the research framework was parsed down. In Section 3, the research results were presented by comparing multiple embedding techniques. In the last section, the findings were concluded and possible forthcoming works in this field were also suggested.

2. METHODOLOGY

2.1. OVERVIEW

Single nucleotide polymorphisms (SNPs) are genetic variations that occur in each living thing on a single nucleotide block in DNA sequences. SNP takes place in various nucleotide blocks scattered throughout the DNA. SNPs data in nucleotide base pairing on rice genomic data is used in this study. The data is nucleotide base pairs that form alleles in the DNA sequence. Therefore, the data must first be converted into an embedding format. In this paper, a comparative study was proposed to assess the performance of different distributed representation methods, i.e., Embedder, Node2Vec, Struc2Vec, and LINE, using 1,232 Indonesian rice SNPs.

2.2. DATA PREPROCESSING

The nucleotide base pair in SNPs data needs to be encoded first. Table 1 shows the encoding method used on the SNPs data.

The data encoding SNPs carried out in this study resulted in three classes or categories of SNPs, namely homozygous major or reference, heterozygous, and homozygous minor or alternate with

encoding 0, 1, and 2, respectively. The encoding process was carried out by considering the presence of alternate alleles in the data obtained. If there is no alternate allele in the data, then the data is categorized as homozygous major and encoded to 0. If in a data there are alternate alleles in allele 1 and allele 2, then the data is categorized as heterozygous and encoded into 1. If both alleles are alternate, then the data is categorized as homozygous minor and encoded into 2. This transforms the biological writing of SNPs data into an embedding format. This encoding process resulted in *n* 0-1-2 SNP sequences, where *n* is the number of samples in the dataset.

TABLE 1. SNPs Data Encoding

ID	[Reference / Alternate] Allele	Allele 1	Allele 2	Encoding	SNP Type
rs1	[A / **T**]	A	A	0	Homozygous major/reference
rs2	[G / **C**]	G	**C**	1	Heterozygous
rs3	[C / **T**]	**T**	C	1	Heterozygous
rs4	[A / **G**]	**G**	**G**	2	Homozygous minor/alternate

2.3. EMBEDDER

Embedder was used to implement entity embedding in this research. Embedder as the first model was based on Guo and Berkhahn's publication which maps categorical variables into Euclidean spaces by using a neural network to learn the entity embeddings. The Embedder has advantages such as less memory usage and faster neural network training, and it could show the intrinsic properties of the categorical variables [13]. Four steps are done inside the Embedder method. First and foremost, the embedder determines the categorical variables by examining the number of unique categories and the data types in each variable in the SNP data. Secondly, the Embedder chooses the smallest of the maximum dimensionality allowed that is supplied from the parameter max_dim to prepare the input SNP data to be embedded. Then, the Embedder uses integer encoding to the dictionary from the second step to prepare them for the next step. The final step of

Embedder is to train a feedforward neural network with two hidden layers on the already prepared rice genomics' SNP data.

In this study, the Embedder was provided with two columns of data from the rice genomics' SNP data which are (1) rice yield which contains a numerical value from each rice sample, and (2) SNPs which contain the SNPs sequences. Before using the Embedder, the data was pre-processed by splitting them into three sets of data, namely training data, validation data, and testing data with the ratio of 70%, 15%, and 15%, respectively. Each of these sets of data was given to the same series of further preprocessing methods. Meanwhile, as the fourth step of the Embedder suggest that the embedding dimension must be provided, this study set two the values of max_dim, i.e., 10 and 50, for hyperparameter tuning.

2.4. COMPARATIVE STUDY

Three other methods were also used to be compared with Embedder, namely Node2Vec, Struc2Vec, and Large-scale Information Network Embedding (LINE) embedding models. Node2Vec, Struc2Vec, and LINE use the same concept as Graph Embedding (GE). The Graph Embedding technique makes use of graphs as input and produces embedding as output. There are four types of input in the GE, i.e., Homogeneous Graph, Heterogeneous Graph, Graph with Auxiliary Information, as well as Graph Constructed from Non-relational Data. Meanwhile, there are also four types of output in the GE, i.e., Node Embedding, Edge Embedding, Hybrid Embedding, and Whole-Graph Embedding [28].

To produce the embedding from the graph data input, Cai et al. [28] mentioned five techniques that can be utilized, such as Matrix Factorization, Deep Learning, Edge Reconstruction, Graph Kernel, and Generative Model. According to Cai et al. [28], GE is used to solve graph analytics problems such as high computation and space cost by converting graph data into a low dimensional space yet still maintaining the graph data's structural information and properties.

Node2Vec, Struc2Vec, and LINE accept a process of preprocessing. Since the method used intends to represent the SNP data in the form of a graph consisting of vertices and edges, a proper criterion for building the edges should be determined. Therefore, a correlation calculation between

each SNP sequence was performed. The correlation calculation produced correlation scores in float data type indicating the degree of correlation between one SNP sequence and another SNP sequence. These correlation scores were then used to determine which vertices (SNP sequence of the n-sample node) should possess an edge over other vertices. A threshold value was given to filter correlation scores between SNP sequences and create graph edges using correlation scores of SNPs that are above the threshold value. Hence, a graph representation of the rice genomic SNPs was produced. This graph is then given to the embedding model and evaluated using a multi-layer perceptron (MLP) regressor.

2.5. EVALUATION METRIC

The Mean Squared Error (MSE) and R2 or R-squared Score were chosen in this research to provide the means to compare and evaluate the models. MSE is a metric that calculates the average squared error of predictions [29]. On the other hand, before averaging the numbers, it computes the square of the difference between the expected and actual values [29]. R-Squared (R2) is a statistical metric used to assess the performance of regression models [30]. It assesses the strength of the association between the dependent variable and regression models on a simple 0-100% scale [30]. The R2 Score determines the dispersion of data points around the regression line [30].

3. RESULT AND DISCUSSION

The result was divided into two parts. The first part explained the R2 and MSE evaluation metrics by using different hyperparameter tuning. The second part was used to present the prediction and actual value plotting of the Embedder's best hyperparameter tuning configuration.

3.1. MODEL EVALUATION METRIC

There are four hyperparameter tuning configurations which are learning rate = 0.001 and embedding dimension = 10, learning rate = 0.0001 and embedding dimension = 10, learning rate = 0.001 and embedding dimension = 50, and learning rate = 0.0001 and embedding dimension = 50. SNPs sequence data of rice samples used in this study were given in several embedding experiments. This study proposes a method of representing SNP data using entity embedding. The proposed entity embedding method is then compared with several other embedding methods,

namely the LINE, Node2Vec, and Struc2Vec by utilizing two evaluation metrics which are R2 Score and MSE. This study used two hyperparameters which are Adam learning rate and embedding dimension. The Adam learning rate choices are 0.001 and 0.0001. Meanwhile, the embedding dimensions are 10 and 50. The experimental results can be seen in Table 2, Table 3, Table 4, and Table 5.

TABLE 2. Hyperparameter (learning rate = 0.001 and embedding dimension = 10)

Model	R2	MSE
Embedder	0.94	0.24
LINE	-3.13	15.89
Node2Vec	-1.00	7.69
Struc2Vec	-2.33	13.77

TABLE 3. Hyperparameter (learning rate = 0.0001 and embedding dimension = 10)

Model	R2	MSE
Embedder	0.94	0.24
LINE	-3.31	16.56
Node2Vec	-0.49	5.75
Struc2Vec	-2.3	13.73

TABLE 4. Hyperparameter (learning rate = 0.001 and embedding dimension = 50)

Model	R2	MSE
Embedder	0.94	0.24
LINE	-2.04	11.69
Node2Vec	-0.93	7.45
Struc2Vec	-2.04	12.59

TABLE 5. Hyperparameter (learning rate = 0.0001 and embedding dimension = 50)

Model	R2	MSE
Embedder	0.94	0.24
LINE	-3.59	17.67
Node2Vec	-0.48	5.70
Struc2Vec	-1.28	9.44

In the experiment, four hyperparameter tunings have been performed, namely a cross combination between the learning rate values of 0.001 and 0.0001 and the embedding dimension values of 10 and 50. From Table 1 to 4, the result displayed that the hyperparameter tuning in Embedder shows none to little performance impact, unlike the comparative method such as LINE, Node2Vec, and Struc2Vec. Experiments showed that the best hyperparameter tuning is achieved by configuring a learning rate parameter of 0.0001 and an embedding dimension of 10. This is because the evaluation metrics R2 and MSE values generated by the Embedder model in that configuration are the best results, with the value of 0.9368 and 0.2425 respectively. However, after analyzing the experimental results, the hyperparameter tuning configuration does not provide a significant difference.

Through the analysis of the experimental results, it can be concluded that Embedder outperformed the comparative model, namely LINE, Struc2Vec, and Node2Vec. The LINE comparative study could not achieve its best result because the LINE method needs millions of nodes in a network to perform well according to Tang *et al.* [15], in which our SNP graph only has 480 nodes for training and 207 nodes for testing. Meanwhile, the other comparative study methods, namely Struc2Vec and Node2Vec, could achieve better evaluation performances than LINE but still fall behind when compared to Embedder. Struc2Vec and Node2Vec's performances are slightly better than LINE's performance. This is due to the random walk algorithm integrated into both models, which reduces the space and time complexity in processing graph-represented data. The difference between Struc2Vec and Node2Vec lies in the graph processing method and perspective. The Struc2Vec model is highly dependent on the structure of the graph-represented

data since it calculates or considers the relation between nodes by considering the symmetry point of the graph. Therefore, this model cannot perform well in processing complex or not-well-structured graph data. This makes the Struc2Vec is also cannot outperform the LINE embedding model in certain hyperparameter tuning.

Neither Node2Vec nor Struc2Vec can outperform the Embedder. The reason is for the embedding in graph distributed representation, the graph needs to use a big dataset such as BlogCatalog, Flickr, and YouTube which is used by Perozzi *et al.* [15] and Wang *et al.* [18]. Hence, the proposed method of Embedder achieved the best result because the MSE comparative models' values are more than 0. Node2Vec is placed in the second rank since it achieved better MSE scores when compared to the other two embedding models, namely the Struc2Vec and LINE. It achieved the best value of 4.54 MSE score which was successfully obtained on the configuration of the learning rate parameter of 0.001 and the embedding dimension of 10. On the other hand, Embedder is constantly producing an MSE value below 0.

4. CONCLUSION

In this paper, SNP embedding has been carried out using entity embedding. The collected data used in this study consists of 687 rice samples taken from four different rice fields with a total of 1,232 SNPs. The SNP data used is tabular data that has been pre-processed so that the writing of the SNP data previously using the biological writing format has been converted into an encoded format. Through the experiments that have been carried out, it has been obtained that the entity embedding using Embedder can outperform the comparative methods used in this study, namely Node2Vec, Struc2Vec, and LINE. From the result, the Embedder entity embedding method achieved the best R2 and MSE scores of 0.9368 and 0.2425, respectively, by using the learning rate configuration of 0.0001 and the embedding dimension of 10. Future works of this research will be conducted by using different embedding methods to attain better SNP data representation performance.

ACKNOWLEDGMENTS

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTERESTS

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

REFERENCES

[1] 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.

[2] C. McMahan, J. Baurley, W. Bridges, et al. A Bayesian hierarchical model for identifying significant polygenic effects while controlling for confounding and repeated measures, Stat. Appl. Genetics Mol. Biol. 16 (2017), 407-419. https://doi.org/10.1515/sagmb-2017-0044.

[3] N. Dominic, T.W. Cenggoro, A. Budiarto, et al. Deep polygenic neural network for predicting and identifying yield-associated genes in Indonesian rice accessions, Sci. Rep. 12 (2022), 13823. https://doi.org/10.1038/s41598-022-16075-9.

[4] N. Dominic, T.W. Cenggoro, B. Pardamean, Systematic literature review on statistics and machine learning predictive models for rice phenotypes, in: 2021 International Conference on Computer Science and Computational Intelligence, 2021.

[5] N. Dominic, T.W. Cenggoro, B. Pardamean, Systematic literature review: Accelerate the rice production for global food security, in: AIP Conf. Proc. 2594 (2023), 080001. https://doi.org/10.1063/5.0109203.

[6] J.W. Baurley, B. Pardamean, A.S. Perbangsa, et al. A bioinformatics workflow for genetic association studies of traits in Indonesian rice, in: Information and Communication Technology, 2014, pp. 356–364. Lecture Notes in Computer Science, 8407 LNCS. https://doi.org/10.1007/978-3-642-55032-4_35.

[7] 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.

[8] J.W. Baurley, A.S. 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), 33–41. https://doi.org/10.14257/ijbsbt.2013.5.6.04.

[9] B. Pardamean, J.W. Baurley, A.S. Perbangsa, et al. Information technology infrastructure for agriculture genotyping studies, J. Inform. Process. Syst. 14 (2018), 655–665. https://doi.org/10.3745/JIPS.01.0029.

[10] J.T. Hancock, T.M. Khoshgoftaar, Survey on categorical data for neural networks, J. Big Data. 7 (2020), 28.

https://doi.org/10.1186/s40537-020-00305-w.

[11] Q. Liu, M.J. Kusner, P. Blunsom, A survey on contextual embeddings, preprint. (2020). https://doi.org/10.48550/arXiv.2003.07278.

[12] P.K. Koo, M. Ploenzke, Improving representations of genomic sequence motifs in convolutional networks with exponential activations, Nat. Mach. Intell. 3 (2021), 258–266. https://doi.org/10.1038/s42256-020-00291-x.

[13] C. Guo, F. Berkhahn, Entity embeddings of categorical variables, preprint. (2016). http://arxiv.org/abs/1604.06737.

[14] F. Hou, R. Wang, J. He, et al. Improving entity linking through semantic reinforced entity embeddings, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020: pp. 6843–6848. https://doi.org/10.18653/v1/2020.acl-main.612.

[15] D. Mezzogori, F. Zammori, An entity embeddings deep learning approach for demand forecast of highly differentiated products, Procedia Manuf. 39 (2019), 1793–1800. https://doi.org/10.1016/j.promfg.2020.01.260.

[16] Y. Ma, Z. Zhang, Travel mode choice prediction using deep neural networks with entity embeddings, IEEE Access. 8 (2020), 64959–64970. https://doi.org/10.1109/ACCESS.2020.2985542.

[17] I. Amihai, M. Chioua, R. Gitzel, et al. Modeling machine health using gated recurrent units with entity embeddings and K-means clustering, in: 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), IEEE, Porto, 2018: pp. 212–217. https://doi.org/10.1109/INDIN.2018.8472065.

[18] B. Perozzi, R. Al-Rfou, S. Skiena, DeepWalk: online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York USA, 2014: pp. 701–710. https://doi.org/10.1145/2623330.2623732.

[19] J. Tang, M. Qu, M. Wang, et al. LINE: Large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Florence Italy, 2015: pp. 1067–1077. https://doi.org/10.1145/2736277.2741093.

[20] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016: pp. 855–864. https://doi.org/10.1145/2939672.2939754.

[21] D. Wang, P. Cui, W. Zhu, Structural deep network embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016: pp. 1225–1234. https://doi.org/10.1145/2939672.2939753.

[22] L.F.R. Ribeiro, P.H.P. Saverese, D.R. Figueiredo, struc2vec: Learning node representations from structural identity, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Halifax NS Canada, 2017: pp. 385–394. https://doi.org/10.1145/3097983.3098061.

[23] G. Partel, C. Wählby, Spage2vec: Unsupervised representation of localized spatial gene expression signatures,

FEBS J. 288 (2020), 1859–1870. https://doi.org/10.1111/febs.15572.

[24] J. Choi, I. Oh, S. Seo, et al. G2Vec: Distributed gene representations for identification of cancer prognostic genes, Sci Rep. 8 (2018), 13729. https://doi.org/10.1038/s41598-018-32180-0.

[25] Q. Zou, P. Xing, L. Wei, B. Liu, Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA, RNA. 25 (2018), 205–218. https://doi.org/10.1261/rna.069112.118.

[26] J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, D. Zhi, Gene2vec: distributed representation of genes based on co-expression, BMC Genomics. 20 (2019), 82. https://doi.org/10.1186/s12864-018-5370-x.

[27] L. Zhang, S. Zhang, K. Balog, Table2Vec: Neural word and entity embeddings for table population and retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Paris France, 2019: pp. 1029–1032. https://doi.org/10.1145/3331184.3331333.

[28] H. Cai, V.W. Zheng, K.C.-C. Chang, A comprehensive survey of graph embedding: problems, techniques, and applications, IEEE Trans. Knowl. Data Eng. 30 (2018), 1616–1637. https://doi.org/10.1109/TKDE.2018.2807452.

[29] K. Palanivel, C. Surianarayanan, An approach for prediction of crop yield using machine learning and big data techniques, Int. J. Computer Eng. Technol. 10 (2019), 110–118.

[30] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Computer Sci. 7 (2021), e623. https://doi.org/10.7717/peerj-cs.623.