Journal of Information Technology and Computer Science
Volume 7, Number 1, April 2022, pp. 110-116
Journal Homepage: www.jitecs.ub.ac.id

Cancer Classification Based on the Features of Itemset Sequence Pattern of TP53 Protein Code Using Deep Miden - KNN

Marji[1], Imam Cholissodin[2], Dian Eka Ratnawati[3], Edy Santoso[4], Nurul Hidayat[5]
[1,2,3,4,5,6]Brawijaya University, Malang, Indonesia
*Corresponding author

Received: 16 April 2022; Accepted: 30 April 2022

Abstract. Cancer is a disease that is still difficult to identify up to today. One of the causes of cancer is genetic modification that because of mutations in p53 gene. Healthy cells have a p53 wild type protein (normal) that is able to manage DNA separation. If DNA mutates, it will be difficult to detect cancer because the composition of the protein has changed. Bioinformatics is a combination of biology and information engineering (TI) that is utilized to manage data. One of the applications of data mining in bioinformatics is the development of pharmaceutical and medical industries. Data mining classification can use variety of methods including K-Nearest Neighbor (KNN), C45, ID3, and several other methods. One of the most reliable data classification methods is KNN. In this study, the development used two algorithms. The first was with the modification of the k-fold method, which divided two data into training data and test data, in which test-1 data and test-2 data were made into slices. The second was by a method for selecting an itemset sequence pattern that had the largest Gain Information, either 2 itemsets, 3 itemsets, and so on (Deep Miden). The best accuracy result of 96.00% was obtained through the process of computation testing in the server based on variations in terms of the number of patterns of Deep Miden itemset sequences and several k values on KNN classification method.

Keywords: feature reduction, dna, cancer disease, gene, p53, deep miden, knn

1 Introduction

Based on the studies regarding the P53 gene, such as the one performed by Ria Kurniarti (2013) the accuracy of 80.00% was obtained by applying the K-Means for clustering and K-Nearest Neighbor (KNN) for classification [1]. Another study conducted by Arinta (2016) obtained an accuracy of 52.57% through applying the modified KNN classification method [2]. Meanwhile, a research conducted by Tawang (2017) obtained an accuracy of 79.17% by applying Naïve Bayes (NB) algorithm [3]. Cluster quality obtained from the research conducted by Laily Putri (2017) by applying the K-Medoids was 77.00% [4]. In another research conducted by Aldy Satria (2018), the accuracy of 80,666% was obtained by applying the NWKNN classification method while the research conducted by Tahtri Nadia Utami (2018) obtained an accuracy of 55.33% by applying the Fuzzy KNN [5][6].

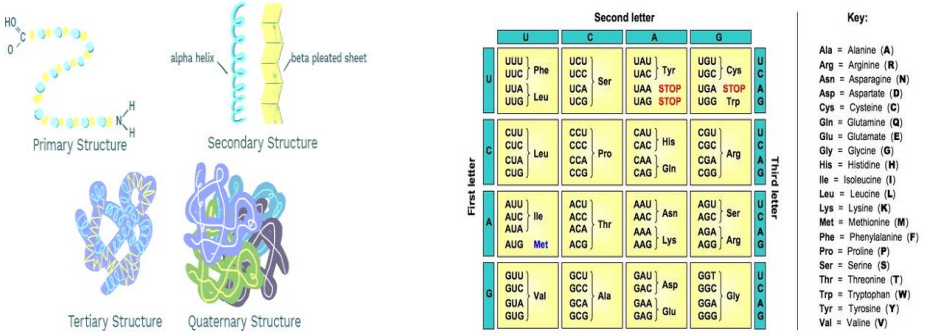
In the research described above, several used 393 features with a P53 mutation code length that employed numerical techniques for the computation process. The way to

convert the char code to numeric code is to utilize a PAM table or use the chance of the code compiler characters appearance. The high or low number of the feature codes used will affect the length of computation time. If fewer features are used, it's expected that the computation will more fast and be able to get deeper patterns of sequences. Good accuracy results can help the many medical staff to do diagnose the patient's condition, whether he is still in a normal condition or has been exposed to specific cancers. This study of research is in line with national and international research or in home base of UB's research in the field of health. The outcome of the purpose research is create k-fold enhancement by modification algorithm, which is an algorithm to obtain the itemset sequence patterns according to the largest Information Gain, and information about cancer in a patient in the form of software. The achievement of this research is to accelerate the determination of the best features of the selected training data according to the non-mutually exclusive-based k-fold modification. The non-mutually exclusive-based k-fold modification was built using itemset sequenced pattern techniques that was modified in accordance with the largest Information Gain value (Deep Miden), then combined with the KNN classification algorithm to obtain the best results from cancer detection.

2 Method

2.1 Protein Structure And Cancer

There are four (4) protein structures based on their shape, i.e. primary structure, secondary structure, tertiary structure, and quarter structure [7]. Protein has the definition that are consist complex molecules like few simple chain blocks, namely amino acids. One of the few benefits of protein is taking care of the adjustment of cell function and carrying out many tasks related to life. The following is a picture of the structure of the protein displayed in Figure 1(i).



(i) (ii)

Figure 1. Protein Structure and Genetic Code [8][9]

Figure 1 (ii) shows the nucleotides contained in DNA with the characters of A, G, T, C. These letters will form a codon, which is a three-letter code. The assembly of codons will be formed a genetic code [10]. The formation of orgasmic tissue by cells employs gene codes for various products. The product is a protein that functions to carry signals, transport molecules such as O_2 and be able to manage cell processes, such as guarding process [11]. Science on DNA series enable the primary form of polypeptides to be recapitulate. Perseverance of DNA series only needed little quantity of DNA because it is easy to produce hundreds of nucleotide sequences. It's describes

the order in which amino acids are mixed to the polypeptide chain that has just been synthesized in the ribosome [10]. Proteins made from genes begin with the process of RNA polymerase connect to chromosomes also identify the beginning point it. There are atoms group (molecules) that opens the multi helix to show DNA chain that form genes, and complement gene duplication to load. The gene operation of duplication become to mRNA is namely with term transcription, while the procedure for change mRNA to protein is using term translation [11].

The cause of cancer is influenced by local factors, including the availability of genes that lead a assignment in cycle of cell has make the beam navel in relationship to the tumor growth cultivate. The deuce cluster of genes are the first cluster as main cause of tumors that are often called oncogenes tumor, like the racial and the c-myc gene. The secondary cluster of genes like tumor suppression cluster that is often namely with term "tumor suppressor gene", for example such as the gene of p53 and the gene of Rb. Meanwhile, most of researchers say that amount 50% of cancers are caused by mutation on these genes [12]. The epigenetic model of cancer has largely lost its appeal because of the mutant array ruling of genes is to be discovered in the humans cells tumor. Therefore, in our purpose focus is to increasingly change over genes and especifically to the genome of the cancer cells.

2.2 1st Propose: Non Mutual Exclusive Of K-Fold Cross Validation

Prediction model is built with the aim of predicting target value data accurately. Therefore, the rate of prediction model error needs to be estimated to measure whether the prediction model that has been built can predict the test data based on the training data used. The easiest and most widely used method of error rate estimation is cross-validation as shown in Figure 2. If the available data is sufficient, the validation set can be organized and can be used to determine the performance of the prediction model. What often happens, however, is that the data used is difficult to obtain. The thing to consider for overcoming this is k-fold cross validation by utilizing part of the available data to train the model and using other parts for testing [13]. In this K-Fold, we propose K-Fold which is non-mutually exclusive (or can be referred to as Overlapping K-Fold), meaning that the id data chosen randomly in each cell may repeat or be in the same cell, or in different cells. The reason behind this is that in real condition, when entered naturally, the patient's data can coincidentally have the same or different values.

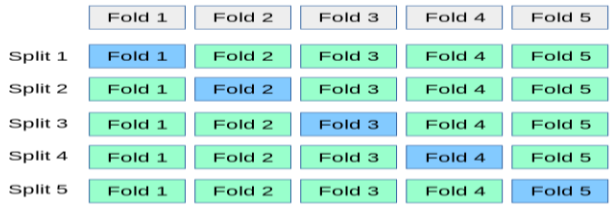


Figure 2. Overlapping K-fold Cross-Validation Process [14]

2.3 2nd Propose: Hybrid Deep Miden And KNN

This study recommends a new representation of the P53 protein code generated from the Deep Miden Algorithm. There were 394 codes in the original data, which were utilized to attribute to different research. This research aims to determine the sequence of support patterns made up of one itemset, two itemsets, and three itemsets. The performance of the three-item calculation will necessitate a long time to compute. As a result, items with high gain values are selected as the features since it is expected to reduce the number of

features in the original data. The steps of the algorithm are listed below:

1. Read the dataset
   Read the protein codes and categories, for example:
   VELWKLLPMEEPPPLSQETFSDQSDPS 0
2. Initializing the pattern
   The initialization is aimed to ascertain the length of a pattern pair of the P53 protein code, which is composed of 20 letters consisting of "ACDEFGHIKLMNPQRSTVWY" which can be used repeatedly. A lengthy sequence arrangement of three, as an instance, like as "CCC", "BCC", etc.
3. While (index of the pattern isn't null)
4. Calculating the pattern supports for the entire data (763 data)
   The length of one protein code is 394. The support patterns such as CCC on the entire data including both paired and unpaired arrangements are determined by this process.
5. Calculating the Information Gain (IG) of a pattern
   If the calculation of one pattern support in the entire data has been carried out, then each value that meets the support is assigned to the pertinent class. Therefore, the support value converts into the class value (1,2,3,..., number of classes). This is useful to make the IG computation easier.
6. Determining the next pattern
   This is the process to establish the next pattern (if one exists), as an instance, the next pattern after the "AAA" pattern is "AAC," and there are no more patterns after the "YYY" pattern.
7. Sorting the patterns in accordance with the Information Gain (IG) obtained
   Even if the classes are distinct, the use of IG features that are categorized descendingly does not guarantee that all the features will have different values. There are still inconsistencies in the data collected from the 50 features, i.e., 40 identical data located in different classes. N, G, VM, V, NR, CL, I, GN, ING, F, YR, VG, VH, IC, C, FQ, VQL are included in the 17 features where changes do not occur; thus, the number of significant features is 33.
8. Displaying the data in a table
   This process records the data in a tabular form. The table contains the names of the attributes with the highest value of Information Gain written ascendingly. As a result, the sequence of attributes consists of S, L, CH, MC, etc. Next, take the five most important features and divide each class into five (5) notes like the samples which are displayed in Table 1.

Table 1. Patterns with the largest Information Gain (IG)

No	<S>	<L>	<CH>	<MC>	<R>	<Class>
<1>	<38>	<32>	<18>	<15>	<26>	<0>
<2>	<38>	<32>	<19>	<16>	<25>	<0>
..	..	..	..	..	..	..
<20>	<38>	<32>	<18>	<17>	<26>	<3>

9. Handling the special case data
   The index values that provide different codes are searched from the two data, and the varied codes can be discovered in several indices. When the various indices are first identified, two patterns of item sets are established. The

difference index, as an instance, is k. The pattern features are then applied to the indices of k and (k + 1) As a result, the new features will have a different value from the two data. The procedures outlined above can be applied to a wide range of data from 753 dissimilar sources. There are 41 features out of 394 that have one different feature value at the minimum.

10. Including the Deep Miden results as a feature for the KNN algorithm
KNN is a method for classifying an object based on training data the distance of which is closest to the object as many as the value of k [15][16]. In general, here are the steps of KNN method.
    a. Establishing the parameter *k* (number of the closest neighbors).
    b. Calculating the term distance of each object against the given test data.
    c. Sorting the objects into groups with the smallest term distance.
    d. Collecting the nearest neighbor classification category as many as *k*.
    e. Predicting the classification results for the calculated instance labels by using the majority of the nearest neighbor category.

3 Results and Analysis

A total of 41 features were employed in the experiment. Previous research can confirm that various classes of data can include different data (one feature with a different value at the minimum). The veracity of a method can be improved by having a broad understanding of various data. In addition, the condition of variation within a class must be less than the condition of variation between classes. This has been proven in previous studies. The results of the test accuracy is shown in Fig. 3 and 4, t, showing the max accuracy is 96.00%. These results indicate that the quality of testing obtained is very good by using features that are more simple than the initial features. Then automatically, the computation time required is also faster.

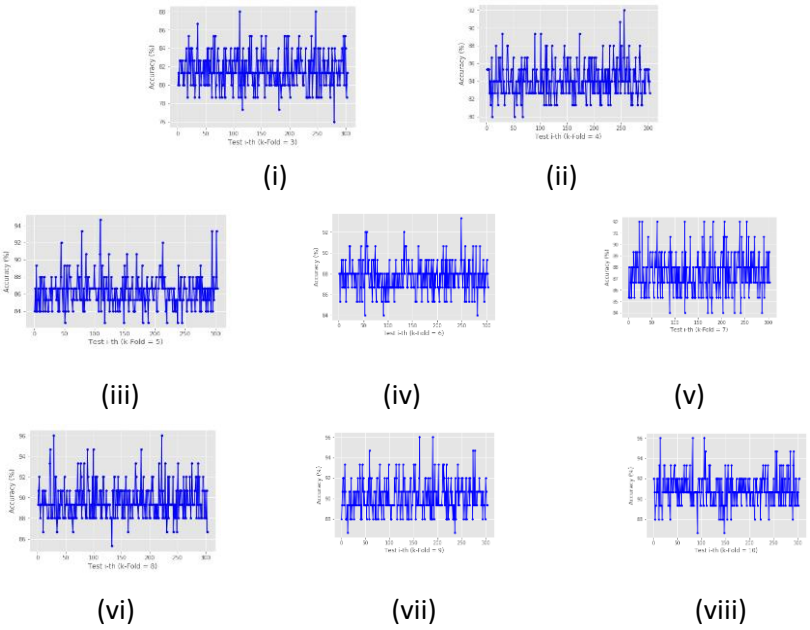


Figure 3. Retrieving data and plotting testing accuracy graph k-Fold 3 to 10

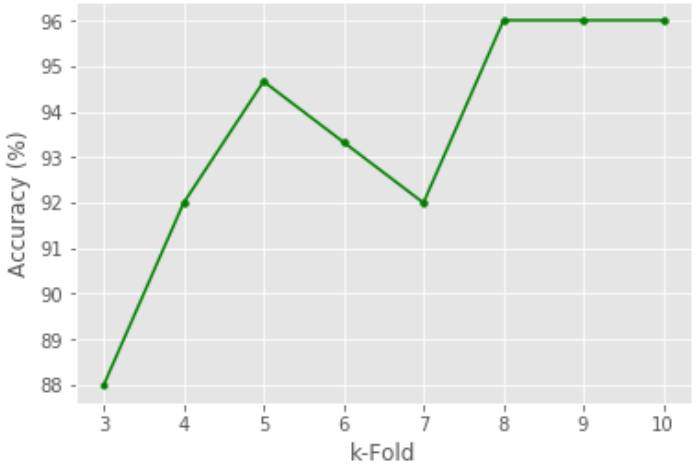


Figure 4. Retrieving max accuracy graph k-Fold 3 to 10

Based on Figure 4, why k = 5 and 6, the performance was degraded. This occurs because the making of the *k*-Fold is random. This also applies to all *k* values. Besides that, there is also no guarantee that the larger or smaller the value of k will always obtain a high accuracy value. The reason is that sufficient pattern bank information is required in the data used as training data despite the small number which is able to represent each pattern vector in the test data. The main topic of discussion is how to discover a non-mutually exclusive pattern with a certain random distribution so that regardless of the value of *k* and the high or small number of pattern banks in the training data, the algorithm used can still provide high accuracy results. However, the best result to use in the final decision is the highest accuracy value from the k-Fold pair between the training data and the test data. In addition, in relation to the adaptive value used in Information Gain, dynamic values should be used. At the end where the data is processed when it is ready, other alternative methods can be used besides KNN so that the fit model built from the training data does not only depend on dominance based on the *k* value and the proximity of the data distance.

4 Conclusion

This study combined the Deep Miden algorithm with KNN which was successfully utilized to diminish data dimensions according to sequential data patterns by: 1. Reading the dataset, 2. Initializing the pattern, 3. While (index of the pattern is still not empty), 4. Calculating the pattern supports for the entire data (763 data), 5. Calculating the Information Gain (IG) of a pattern, 6. Determining the next pattern, 7. Sorting the patterns in accordance with the Information Gain (IG) obtained, 8. Displaying the data in a table, and 9. Handling the case data that is specific to the data of P53 protein that is classified in 394 characters or features. Curtailment of data dimension serves to accelerate the computation time without sacrificing information in the data. According to the variation and grade testing, the veracity value obtained from the curtailment of features from 394 to 41 features is 96.00%. It is proven that the reduction of features can provide good important data patterns with the tendency of faster computation time and quite high accuracy values. Future studies are expected to not only focus on dimension reduction but also be able to reduce the data to choose the best training data, i.e. by using more data than the data used in the research that has

previously been done. In addition, hybrid techniques by adding better classification methods can be employed to obtain more optimal results. The example is using an encoder technique for the process of dimensional reduction that can be proceeded simultaneously to the classification process (Deep Learning) or using the KNN kernel.

References

1. R. Kurnianti. 2013. Penggunaan Metode Pengelompokan K-Means pada klasifikasi KNN untuk penentuan jenis kanker berdasarkan susunan protein. Skripsi PTIIK UB.
2. Retwitasari, A., 2016. Penentuan Jenis Kanker Berdasarkan Struktur Protein Menggunakan Algoritma Modified K-Nearest Neighbor (MKNN). Skripsi PTIIK UB.
3. Wulandari, T. 2018. Classification Of Cancer Types Based On Protein Structure Using The Naive Bayes Algorithm, address http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2718. Skripsi Filkom UB.
4. Rizby, L. P. 2018. Clustering pasien kanker berdasarkan struktur protein dalam tubuh menggunakan metode K-Medoids, alamat http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2740, Skripsi Filkom UB.
5. Satria, A. 2018. Klasifikasi Jenis Kanker Berdasarkan Struktur Protein Menggunakan Metode Neighbor Weighted K-Nearest Neighbor (NWKNN), alamat : http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/ view/4988, Skripsi PTIIK UB.
6. Utami, T. N. 2018. Implementasi Fuzzy k-Nearest Neighbor (Fk-NN) untuk Klasifikasi Jenis Kanker berdasarkan Susunan Protein, address : http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4105, Skripsi PTIIK UB.
7. Wang, J. T., et al. 2006. Data mining in bioinformatic (Advanced information and knowledge processing). Berlin Heidelberg: Springer London.
8. BioNinja, "Transcription and Translation," [online] Available at: < http://www.old-ib.bioninja.com.au/standard-level/topic-3-chemicals-of-life/35-transcription-and-transl.html >. [Accessed January, 29 2020]
9. ThoughtCo, "Learn About the 4 Types of Protein Structure," [online] Available at: < https://www.thoughtco.com/protein-structure-373563 >, 2019. [Accessed Jan, 29 2020]
10. Murray, R. K., Granner, D. K., and Rodwell, V. W. 2006. Harper's Illustrated Biochemistry (27 ed.). The McGraw-Hill Companies inc.
11. Keedwell, E., and Narayanan, A. 2005. Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems. Hoboken, New Jersey: John Wiley & Sons, Inc.
12. Pusztai, L., Lewis, C., and Yap, E. 1996. Cell Proliferation in Cancer- Regulation Mechanisms of Neoplastic Cell Growth. Oxford: Oxford University Press.
13. Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning Second, New York: Springer-Verlag.
14. scikit-learn, "Cross-validation: evaluating estimator performance," [online] Available at: < https://scikit-learn.org/stable/modules/cross_validation.html >, 2007 - 2019. [Accessed Jan, 29 2020]
15. Baharsyah, I., Cholissodin, I., and Setiawan, B. D. 2014. Klasifikasi Deep Sentiment Analysis E-Complaint Universitas Brawijaya Menggunakan Metode K-Nearest Neighbor," in Journal PTIIK Doro, 2014. Doro 2014. Vol. 3 no. 8.
16. Afandie, M. N., Cholissodin, I., and Supianto, A. A. 2014. Implementasi Metode K-Nearest Neighbor Untuk Pendukung Keputusan Pemilihan Menu Makanan Sehat Dan Bergizi in Journal PTIIK Doro, 2014. Doro 2014. Vol. 3 no. 1.