proceedings

MDPI

Proceedings

Optimization for Gene Selection and Cancer Classification †

Hülya Başeğmez [1,*], Emrah Sezer [2] and Çiğdem Selçukcan Erol [2]

[1] Management Information Systems Department, Beykent University, 34398 Istanbul, Turkey
[2] Informatics Department, Istanbul University, 34134 Istanbul, Turkey; emrahsezer39@gmail.com (E.S.); cigdems@istanbul.edu.tr (C.S.E.)
* Correspondence: hulyasezer@beykent.edu.tr or hulyabasegmez@gmail.com
† Presented at the 7th International Management Information Systems Conference, Online, 9–11 December 2020.

Abstract: Recently, gene selection has played an important role in cancer diagnosis and classification. In this study, it was studied to select high descriptive genes for use in cancer diagnosis in order to develop a classification analysis for cancer diagnosis using microarray data. For this purpose, comparative analysis and intersections of six different methods obtained by using two feature selection algorithms and three search algorithms are presented. As a result of the six different feature subset selection methods applied, it was seen that instead of 15,155 genes, 24 genes should be focused. In this case, cancer diagnosis may be possible using 24 candidate genes that have been reduced, rather than similar studies involving larger features. However, in order to see the diagnostic success of diagnoses made using these candidate genes, they should be examined in a wet laboratory.

Keywords: feature selection; classification; ovarian cancer; Consistency Based Feature Selection; Correlation Based Feature Selection; Genetic Search; Best First; Rank Search; Gain Ratio Based Feature Selection

Citation: Başeğmez, H.; Sezer, E.; Selçukcan Erol, Ç. Optimization for Gene Selection and Cancer Classification. *Proceedings* **2021**, *74*, 21. https://doi.org/10.3390/proceedings2021074021

Published: 16 March 2021

1. Introduction

The DNA microarray enables us to understand the structure of many genes that provide information about the physiological processes and disease etiology mediated by these genes. Regulation of a gene expression occurs during the adaptation of DNA to reporter ribonucleic acid (mRNA). DNA microarrays are a tool used for the identification and measurement of mRNA transcripts found in cells [1].

Microarray datasets play an important role in cancer detection. However, the large size of these datasets makes classification difficult due to the presence of many irrelevant and unnecessary features. For this reason, feature (gene) selection has a very important place in this field thanks to its ability to remove features that are not required from the existing structure.

The microarray gene expression dataset contains information on the expression levels of genes in the particular tissue and cell. These data are used as a key source of information in different biological studies and analyzes. Therefore, microarray data are very useful in the field of tumor and cancerous gene detection.

Microarray data generally include expression profiles of genes for both cancerous (tumor) and non-cancerous (normal) cells. Proper analysis will help the medical doctor and drug designer identify the genes responsible for cancers and take action before the disease becomes incurable. Therefore, microarray gene expression data are important because treatment becomes easier after detection [2].

Generally, the microarray data contain a small number of samples (around 100) and a large number of features (approximately 6000 to 60,000) that lead to the "curse of dimensionality" [3]. Most features are unnecessary and/or irrelevant in such data. Because

expression values can indicate the occurrence of cancer, the features that are most relevant are called biomarkers. Hence, finding biomarkers is an important research problem. Irrelevant features increase the accuracy and calculation time of the cancer detection system. In short, all features (genes) are not responsible for cancer, only a very small fraction of the total number of genes cause cancer. This, in turn, expresses the importance of the choice of features that eliminate irrelevant and/or unnecessary data in the dataset and make detection faster and more accurate [2,4].

Different feature selection methods in the literature determine the optimal features differently. Therefore, different results occur when different methods are applied one by one. If we apply a number of methods separately and take the combination or intersection of the results we get from these methods, we not only get the most important information from all methods, but also increase our chances of improvement in the prediction performance of the system. Therefore, the purpose of combining multiple feature selection methods is to increase the maximum accuracy achieved with a single method. Because, the combination can overcome the errors of other methods in different parts of the input field while increasing accuracy by providing complementary views on the importance of features [2].

For this purpose, in this study, subsets of features were selected using six different methods obtained by using two feature selection algorithms, and three search algorithms, and classification studies with them were performed and the results were examined. In addition, the intersections of these six different feature subsets were also examined. In this study, the Ovarian cancer dataset produced as a result of a study by Zhu et al. (2007) was used. This dataset is accessible to researchers.

Ovarian cancer is one of the most common gynecological cancers with the highest mortality rate. It is the eighth most common cancer among women in the world and the 18th most common cancer in general [5].

Ovarian cancer arises at advanced clinical stages in more than 80% of patients and is associated with 5-year survival in 35% of this number. In contrast, 5-year survival exceeds 90% for patients with stage I ovarian cancer, and most patients treat their disease with surgery alone. Therefore, increasing the number of women diagnosed with stage I disease is expected to have a direct impact on the mortality and economy of this cancer without the need to change the approaches used in surgery or chemotherapy [6].

The American Cancer Society estimated that in 2020, 21,750 new female cases of ovarian cancer will be detected in the United States, and 13,940 of these cases will die of ovarian cancer. In addition, a woman's lifetime risk of developing ovarian cancer was expressed as approximately 1/78 and the lifetime rate of dying from ovarian cancer as 1/108 [7].

2. Research Methodology

In this section, firstly, the dataset used in the study will be explained. Later, data processing methods and algorithms we use will be discussed in detail.

2.1. Ovarian Dataset Description

The Ovarian cancer dataset (8-7-02) used in this research was produced as a result of a study by Zhu et al. (2007). Researchers can easily access this dataset from Reference [8]. The mentioned dataset consists of 15154 genes (features), 253 observations and 2 classes. The current observation group consists of 162 people with the disease and 91 healthy people. This dataset was produced using the WCX2 protein chip and is very different from the Ovarian cancer dataset (4-3-02).

2.2. Algorithms

In this research, six different methods obtained by using two feature selection algorithms, and three search algorithms were evaluated. The classification algorithm

Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT) were evaluated through the ovarian dataset. These algorithms are summarized briefly in Table 1.

Table 1. Used Algorithms.

Algorithms		Descriptions
Feature Selection Algorithms	Consistency Based FS	Works with the principle of choosing a consistency based feature subset [9]
	Correlation Based FS	Sort features based on a correlation-based evaluation function [10–12]
Search Algorithms	Genetic Search	Performs a search using the simple genetic algorithm described in [13]
	Best First	Heuristic search method that searches the domain of feature subsets with greedy hill climbing enriched with a backtracking facility [14]
	Rank Search	It is a search method that works with the sorter search principle [15]
Classification Algorithms	Support Vector Machine	Perform classification with the help of a linear or nonlinear function [16]
	Random Forest Algorithm	Random forests are a collection of tree-type classifications based on the idea of using a forest for classification purposes [17]
	Decision Tree	It creates a tree-shaped structure in order to make a decision [16,18]

3. Experimental Analysis

In this section, we discuss data preprocessing, classification stage and feature selection studies.

3.1. Data Preprocessing

The Ovarian cancer dataset used in this research was produced as a result of a research by Zhu et al. (2007). The Ovarian dataset was downloaded as an Arff file and the Weka libraries on Python were used to read and study this file. After reading the Arff file, the data was converted to the Dataframe format in the Pandas library and the examination phase was started. Later, the types of features were examined and it was seen that only the feature named 'Class' was categorical and the other features were numeric (See Table 2).

When the Class feature was examined, it was seen that there were a total of 253 observations and two classes. In addition, it was observed that the current observation group consisted of 162 people with the disease and 91 healthy people (See Figure 1).

Table 2. Feature types and number of features in the Ovarian Dataset.

Feature Type	Count
Numeric (Continuous)	15,154
Categorical	1

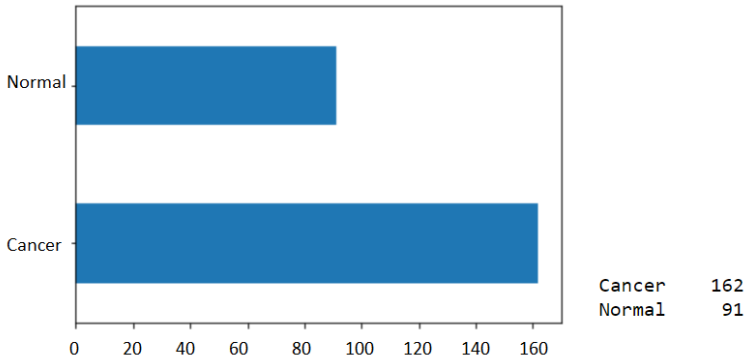

Figure 1. Distribution of current observations in the Ovarian dataset.

When the averages, quarters, minimum and maximum values of the columns were examined, it was observed that all columns were compressed between 0 and 1, but the averages and distributions differed. Descriptive statistics can be seen in Table 3.

Table 3. Descriptive Statistics.

MZ-7.86E-05	MZ2.18E-07	MZ9.60E-05	MZ0.000366014	…	MZ19992.874	MZ19995.513	Class
count	253	253	253	…	253	253	253
unique	NaN	NaN	NaN	…	NaN	NaN	2
top	NaN	NaN	NaN	…	NaN	NaN	Cancer
freq	NaN	NaN	NaN	…	NaN	NaN	162
mean	0.53245	0.462623	0.465859	…	0.430548	0.430548	NaN
std	0.183136	0.196834	0.196418	…	0.155192	0.155192	NaN
min	0	0	0	…	0	0	NaN
25%	0.39785	0.329667	0.321841	…	0.330282	0.330282	NaN
50%	0.537636	0.461537	0.459768	…	0.433102	0.433102	NaN
75%	0.655912	0.593407	0.597701	…	0.541554	0.541554	NaN
max	1	1	1	…	1	1	NaN
11 rows × 15,155 columns							

Extreme values were checked with the box chart. Due to the compression of the data between 0 and 1, it was determined that the samples with the limit values were inconsistent, but these samples were not excluded from the data due to the small number of samples. This process is illustrated in Figure 2. Before the classification study, by applying z-score normalization, a relatively better distribution of the data was attempted.

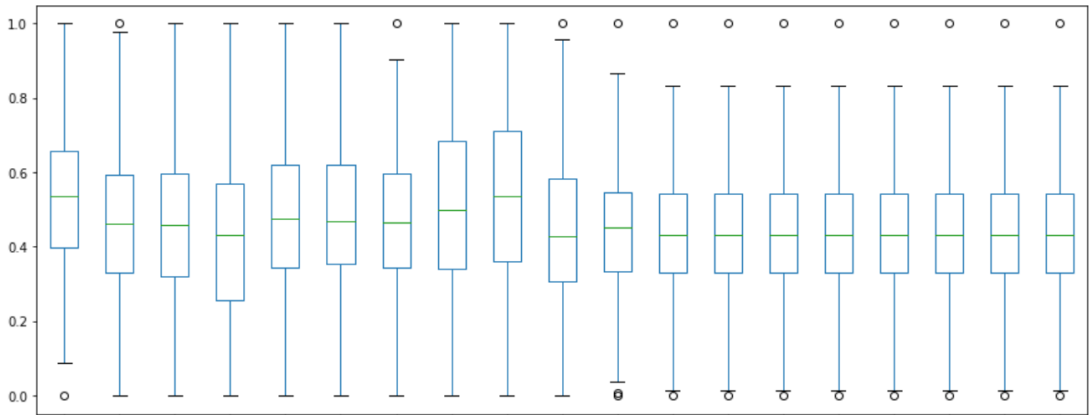

Figure 2. Box Plot.

Later, it was checked whether there were any missing records in all the data and it was seen that there were no missing records.

3.2. Classification Stage

In the classification phase of our study, firstly, the current data were classified using three different algorithms, including five-fold cross validation SVM, RF, and DT. When using RF on the Ovarian dataset, as seen in Table 4, 98.8% classification accuracy and 0.98809 F-score values were reached.

Table 4. Classification Results for Random Forest (RF).

Confusion matrix: $\begin{bmatrix} 162 & 0 \\ 3 & 88 \end{bmatrix}$

Accuracy: 0.988142292490
F-score: 0.98809731937978
Feature Count: 15,155

	Precision	Recall	F1-Score	Support
Cancer	0.98	1	0.99	162
Normal	1	0.97	0.98	91
Micro avg	0.99	0.99	0.99	253
Macro avg	0.99	0.98	0.99	253
Weighted avg	0.99	0.99	0.99	253
(253, 15,154)				

With the classification study using DT on the Ovarian dataset, 95.7% classification accuracy and 0.957 F-score values were achieved, as seen in Table 5. According to the RF algorithm on the lower levels it was observed to obtain the classification performance.

In the case of SVM, as seen in Table 6, 98.8% classification accuracy and 0.98812 F-score values were achieved. It has been observed that these values are very similar to the values obtained by the study performed with the RF Algorithm.

Table 5. Classification Results for Decision Tree (DT).

Confusion matrix: $\begin{bmatrix} 156 & 6 \\ 5 & 86 \end{bmatrix}$

Accuracy: 0.9565217391304
F-score: 0.95657322838352
Feature Count: 15,155

	Precision	Recall	F1-Score	Support
Cancer	0.97	0.96	0.97	162
Normal	0.93	0.95	0.94	91
Micro avg	0.96	0.96	0.96	253
Macro avg	0.95	0.95	0.95	253
Weighted avg	0.96	0.96	0.96	253
(253, 15,154)				

Table 6. Classification Results for Support Vector Machine (SVM).

Confusion matrix: $\begin{bmatrix} 161 & 1 \\ 2 & 89 \end{bmatrix}$

Accuracy: 0.9881422924901
F-score: 0.98812777901896
Feature Count: 15,155

	Precision	Recall	F1-Score	Support
Cancer	0.99	0.99	0.99	162
Normal	0.99	0.98	0.98	91
Micro avg	0.99	0.99	0.99	253
Macro avg	0.99	0.99	0.99	253
Weighted avg	0.99	0.99	0.99	253
(253, 15,154)				

A summary of three different classification studies can be seen in Table 7.

Table 7. Comparison of SVM, DT and RF.

	Data	Classification	ACC	F-Score	Feature Count
0	All	SVM	0.988142	0.988128	15,155
1	All	DT	0.956522	0.956573	15,155
2	All	RF	0.988142	0.988097	15,155

3.3. Feature Selection Studies

The possible feature subset space for the Ovarian dataset was computed as a 4562 digit number, which is the equivalent of 2^{15153}.

On the existing Ovarian dataset, Correlation Based Feature Selection and Consistency Based Feature Selection algorithms and six different feature subsets selected by Cartesian matches of Best First, Genetic Search and Rank Search algorithms were conducted. The GainRatioAttributeEval algorithm, which the Rank Search algorithm uses

as the default algorithm to determine ranking scores, has been preferred. These algorithms are called with the Weka library on Python.

The abbreviations given in Table 8 were created to facilitate analysis. As can be seen from Table 8, six different feature selection applications were made on the Ovarian data and the selected feature numbers were obtained as in Table 9.

Table 8. Used Feature Selection and Search Algorithms.

Abbreviation	Feature Selection Algorithms	Search Algorithms
CfsBest	Correlation Based FS	Best First
CfsGen	Correlation Based FS	Genetic Search
CfsRank	Correlation Based FS	Rank Search
ConBest	Consistency Based FS	Best First
ConGen	Consistency Based FS	Genetic Search
ConRank	Consistency Based FS	Rank Search

Table 9. Number of features selected.

	Best First	RankInfoGain	Genetic
Consistency	3	15	2346
CFS	35	42	3749

When Table 9 is examined, it can be said that the genetic algorithm behaves more greedily because it chooses quite a lot of features compared to the other methods used. The classification studies applied on all data with the data subsets obtained as a result of matching search algorithms and feature selection algorithms were carried out and the values obtained are given in Table 10.

Table 10. Comparison of Classification Results.

	Data	Classification	ACC	F-Score	Feature Count
0	All	SVM	0.988142	0.988128	15.155
1	All	DT	0.956522	0.956573	15.155
2	All	RF	0.988142	0.988097	15.155
3	CfsBest	SVM	1.000000	1.000000	36
4	CfsBest	DT	0.956522	0.956357	36
5	CfsBest	RF	0.992095	0.992075	36
6	CfsGen	SVM	0.992095	0.992095	3750
7	CfsGen	DT	0.968379	0.968301	3750
8	CfsGen	RF	0.988142	0.988097	3750
9	CfsRank	SVM	1.000000	1.000000	43
10	CfsRank	DT	0.976285	0.976226	43
11	CfsRank	RF	0.992095	0.992075	43
12	ConBest	SVM	0.996047	0.996043	4
13	ConBest	DT	0.992095	0.992113	4
14	ConBest	RF	0.996047	0.996043	4
15	ConGen	SVM	0.980237	0.980213	2347
16	ConGen	DT	0.956522	0.956357	2347
17	ConGen	RF	0.964427	0.964195	2347
18	ConRank	SVM	0.980237	0.980213	16
19	ConRank	DT	0.968379	0.968301	16
20	ConRank	RF	0.964427	0.964469	16

Then, grouping was made based on the datasets in Table 10, and Table 11 was obtained by taking the average of accuracy and F-score values. The averages in Table 11 are ranked according to ACC.

When Table 11 is examined, it is seen that the classification studies performed with the data sub-sets obtained by the four different feature selection studies achieved a relatively higher accuracy rate and F-score value than the classification studies performed with all data. When compared with the other two data sub-sets, it was seen that more successful classification studies with an acceptable difference were performed with fewer features than the classification studies conducted with all features. At this stage, it can be said that all feature subset selection studies have produced good results. It has been observed that the approaches using Genetic Search create subsets containing relatively more features than others. In this study, in the applications where the Genetic Search algorithm is used, it can be said that the Genetic Search algorithm has the opportunity to be explained with less features because of the greedy behavior on this data, but it uses more features.

Table 11. The Average of Classification Results for each algorithm.

Data	ACC	F-Score	Feature Count
ConBest	0.994730	0.994733	4
CfsRank	0.989460	0.989434	43
CfsBest	0.982872	0.982811	36
CfsGen	0.982872	0.982831	3750
All	0.977602	0.977599	15.155
ConRank	0.971014	0.970994	16
ConGen	0.967062	0.966922	2347

When the averages of the results of the studies are plotted, it is obvious that very similar results are obtained (See Figure 3). In Figure 3, ACC is shown by blue, F-score is shown by orange. This shows that similar results can be achieved with far fewer features, and it can be said that the application of feature selection is beneficial for this classification study.

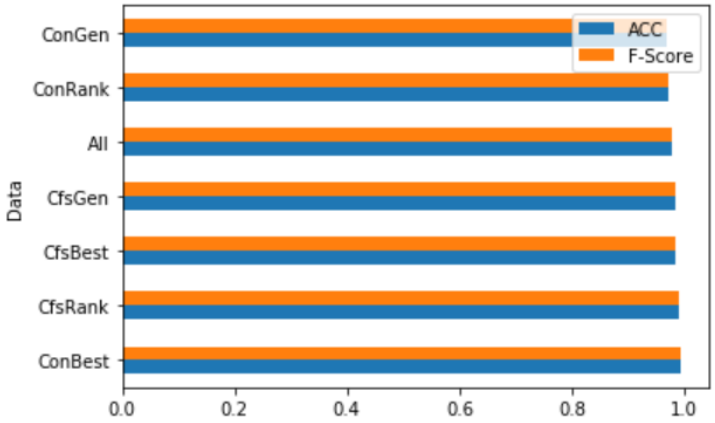

Figure 3. Average of Classification results.

Table 12 was created to examine whether all the selected genes were selected by which applications and their intersections. In Table 12, 0 (zero) means that the relevant gene was not selected, and 1 means that it was selected. It was observed that six different feature selection practices selected a total of 5532 features. Some of these genes have been selected in more than one application. It was calculated in how many different

applications the gene in each row was selected and these calculation results were added as a column in Table 12.

Table 12. Genes, and Applications that Select Genes.

Genes	CfsBest	CfsGen	CfsRank	ConBest	ConGen	ConRank	SumOfRow
MZ0.022435711	1	0	1	0	1	0	3
MZ2.8864971	1	0	0	0	1	0	2
MZ11.165473	1	0	0	0	0	0	1
...	...	...	...	...	...	...	...
MZ19924.315	0	0	0	0	1	0	1
MZ19932.22	0	0	0	0	1	0	1
MZ19971.766	0	0	0	0	1	0	1
5532 rows × 7 columns							

In order to examine the intersections, the graph in Figure 4 was obtained by using the Upset function in the UpSetR library in the R language. In this graph, the sizes of the clusters are shown in the row, which clusters intersect is shown in the points in the middle and the number of elements at the intersections is shown in the graphics and numbers at the top.

At this stage, the intersection table obtained to examine the features in at least three different clusters was filtered and shown in Table 13.

Table 13. Intersection Table.

Genes	CfsBest	CfsGen	CfsRank	ConBest	ConGen	ConRank	SumOfRow
MZ244.66041	1	1	1	0	1	1	5
MZ244.95245	1	1	1	1	0	1	5
MZ674.57738	1	1	1	0	1	0	4
MZ245.53704	1	1	1	0	0	1	4
MZ246.70832	1	0	1	0	1	1	4
MZ417.73207	1	1	1	0	0	1	4
MZ2.8234234	0	1	1	1	1	0	4
MZ435.46452	1	1	1	0	0	1	4
MZ0.022435711	1	0	1	0	1	0	3
MZ434.68588	0	1	1	0	0	1	3
MZ247.00158	0	1	1	0	0	1	3
MZ246.41524	0	1	1	0	0	1	3
MZ245.8296	0	1	1	0	0	1	3
MZ222.41828	0	1	1	0	0	1	3
MZ4906.9617	1	1	1	0	0	0	3
MZ435.07512	1	0	1	0	0	1	3
MZ555.74254	1	1	1	0	0	0	3
MZ194.41064	1	1	0	0	1	0	3
MZ433.90794	1	1	1	0	0	0	3
MZ261.88643	1	1	1	0	0	0	3
MZ246.12233	1	0	1	0	0	1	3
MZ245.24466	1	0	1	0	0	1	3
MZ244.07686	1	1	1	0	0	0	3
MZ435.85411	0	1	1	0	0	1	3

As a result, it can be said that it would be beneficial to consider the genes corresponding to the 24 features in Table 13 in cancer diagnosis.

It can be seen from Figure 4 that CfsGen (Correlation Based FS and Genetic Search) alone selected 3142 genes that other algorithms did not select. Similarly, the intersection of ConGen and CfsGen has alone selected 580 genes.

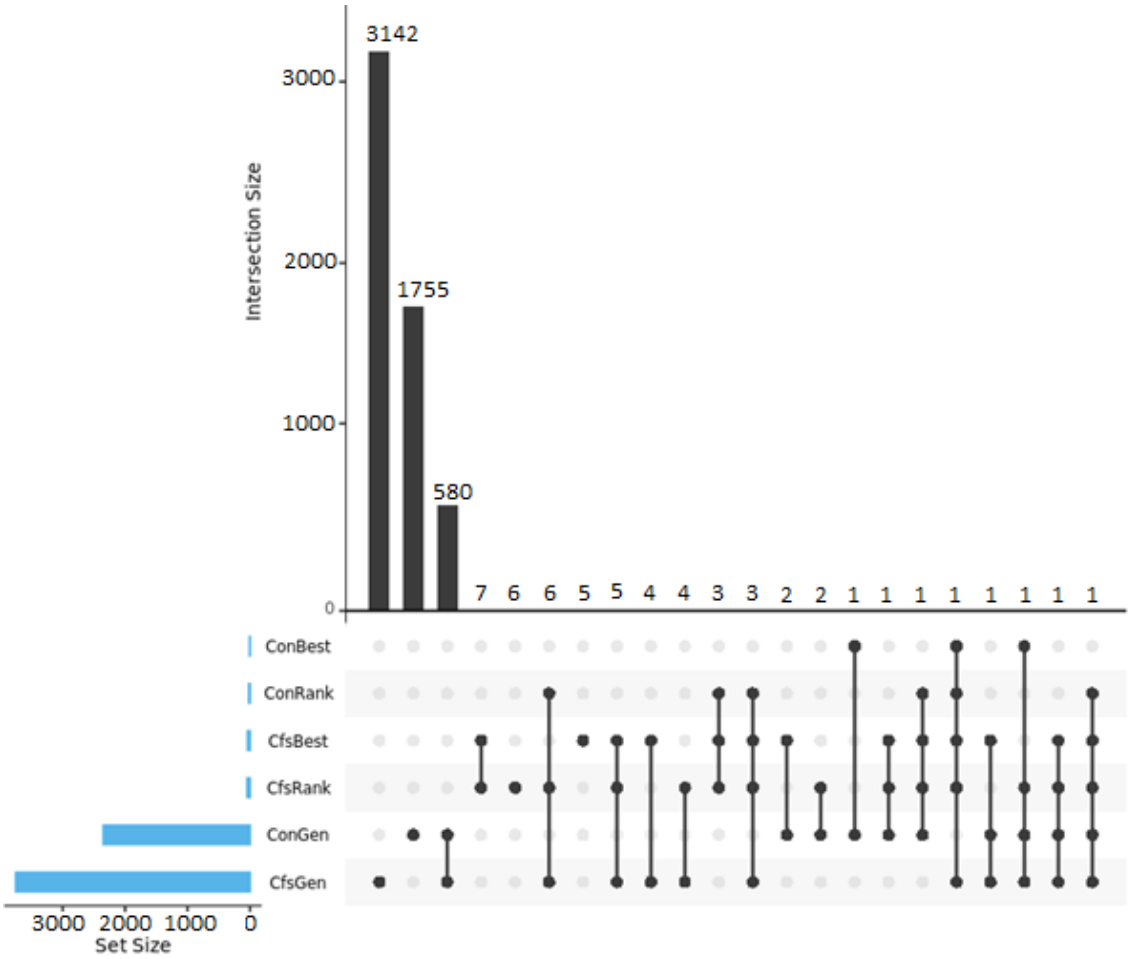

Figure 4. Intersection chart.

4. Discussion and Results

In this study, in order to improve a classification study on cancer diagnosis by using microarray data, the selection of genes with high descriptiveness for use in cancer diagnosis by using feature selection methods was studied. Studies have been conducted in the literature to evaluate the intersections of different feature subsets by selecting them. In this study, intersection sets of variable subsets selected using six different methods were also examined. However, since the size of the data used in the study was not capable of representing the entire human population, the study suggested an approach, and it was not possible for the results to contain certain judgments in terms of genetics.

There are many studies and approaches to gene selection in cancer detection in the literature. Our approach in this study is to examine the selection frequencies of genes selected with different feature selection studies, and the more frequently selected genes may have higher cancer descriptors.

As a result, it has been shown that instead of trying to predict ovarian cancer over 15,155 genes, it can be predicted with 24 genes selected by the majority of the practice of selecting six different feature subsets from among 15,155 genes. Thanks to this reduction in the number of genes, instead of similar studies with larger features, cancer detection may be possible with fewer microarray data, and workforce and cost requirements can be reduced by conducting studies only for the relevant genes in the subsequent diagnostic

stages. In addition, it is thought that higher diagnostic success can be achieved by excluding variables with low explanatory value from the study. However, diagnoses made using these candidate genes need to be examined in a wet laboratory to see diagnostic success.

Within the scope of future studies, it may be possible to make a wider range of evaluation and gene selection by using different feature subset selection methods and classification algorithms.

Author Contributions: In this study, H.B. prepare the model experiments, interpret the result and prepare the manuscript. E.S. contributed to the formal analysis and software. Ç.S.E. is a research advisor and she provide intitutive explanation the manuscripts. All authors have read and agreed to the published version of the manuscript.

Funding: The authors have not received any financial support for this study.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset used in this research was produced as a result of a research by Zhu et al. (2007) as stated in Section 3.1.

Conflicts of Interest: The authors declare no conflict of interest regarding the publication of this paper.

References

1. Özkan, Y.; Erol, Ç. *Biyoenformatik DNA Mikrodizi: Veri Madenciliği*; Papatya Yayıncılık Eğitim: Istanbul, Turkey, 2015.
2. Ghosh, M.; Adhikary, S.; Ghosh, K.K.; Sardar, A.; Begum, S.; Sarkar, R. Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. *Med. Biol. Eng. Comput.* **2019**, *57*, 159–176, doi:10.1007/s11517-018-1874-4.
3. Vaidya, A.R. Neural mechanisms for undoing the "curse of dimensionality". *J. Neurosci.* **2015**, *35*, 12083–12084, doi:10.1523/JNEUROSCI.2428-15.2015.
4. Sezer, E. *An Application on Feature Selection for Classification*; Marmara University Institute of Social Sciences: Istanbul, Turkey, 2018.
5. Momenimovahed, Z.; Tiznobaik, A.; Taheri, S.; Salehiniya, H. Ovarian cancer in the world: Epidemiology and Risk factors. *Int. J. Womens Health* **2019**, *11*, 287–299, doi:10.2147/IJWH.S197604.
6. Petricoin, E.F.; Ardekani, A.M.; Hitt, B.A.; Levine, P.J.; Fusaro, V.A.; vd Steinberg, S.M. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **2002**, *359*, 572–577, doi:10.1016/S0140-6736(02)07746-2.
7. The American Cancer Society. 2020. Available online: https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html (accessed on 13 May 2020).
8. Zhu, Z.; Ong, Y.-S.; Dash, M. Markov Blanket-Embedded Genetic Algorithm for Gene Selection. *Pattern Recognit.* **2007**, *40*, 3236–3248.
9. Liu, H.; Setiono, R. A probabilistic approach to feature selection-a filter solution. In Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, 3–9 September 1996; pp. 319–327.
10. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999. Available online: https://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.pdf (accessed on 2 March 2021).
11. Jungjit, S. New Multi-Label Correlation-Based Feature Selection Methods for Multi-Label Classification and Application in Bioinformatics. Ph.D. Thesis, University of Kent, Canterbury, UK, 2016.
12. Sun, Y.; Wang, F.; Wang, B.; Chen, Q.; Engerer, N.A.; Mi, Z. Correlation Feature Selection and Mutual Information Theory Based Quantitative Research on Meteorological Impact Factors of Module Temperature for Solar Photovoltaic Systems. *Energies* **2017**, *10*, 7, doi:10.3390/en10010007.
13. Goldberg, D.E.; Holland, J.H. Genetic Algorithms and Machine Learning. *Mach. Learn.* **1988**, *3*, 95–99.
14. Weka Class BestFirst. Available online: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/BestFirst.html (accessed on 2 March 2021).
15. Gnanambal, S.; Thangaraj, M.; Meenatchi, V.T.; Gayathri, V. Classification Algorithms with Attribute Selection: An evaluation study using WEKA. *Int. J. Adv. Netw. Appl.* **2018**, *9*, 3640–3644.
16. Balaban, E.; Kartal, E. *Veri Madenciliği ve Makine Öğrenmesi Temel Kavramlar, Algoritmalar, Uygulamalar*, 1 st ed.; Çağlayan Kitap & Yayıncılık & Eğitim: Istanbul, Turkey, 2019.
17. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
18. Brijain, M.; Patel, R.; Kushik, M.; Rana, K. A Survey on Decision Tree Algorithm for Classification. *Int. J. Eng. Dev. Res.* **2014**, *2*, 1–5. Available online: www.ijedr.org (accessed on 2 March 2021).