SECB3203

PROGRAMMING FOR BIOINFORMATICS

GROUP 10

NURATHIRAH BINTI MUHAMAD ZAKI

YUSRA NADATUL ALYEEA BINTI YUSRAMIZAL

OVARIAN DATASET

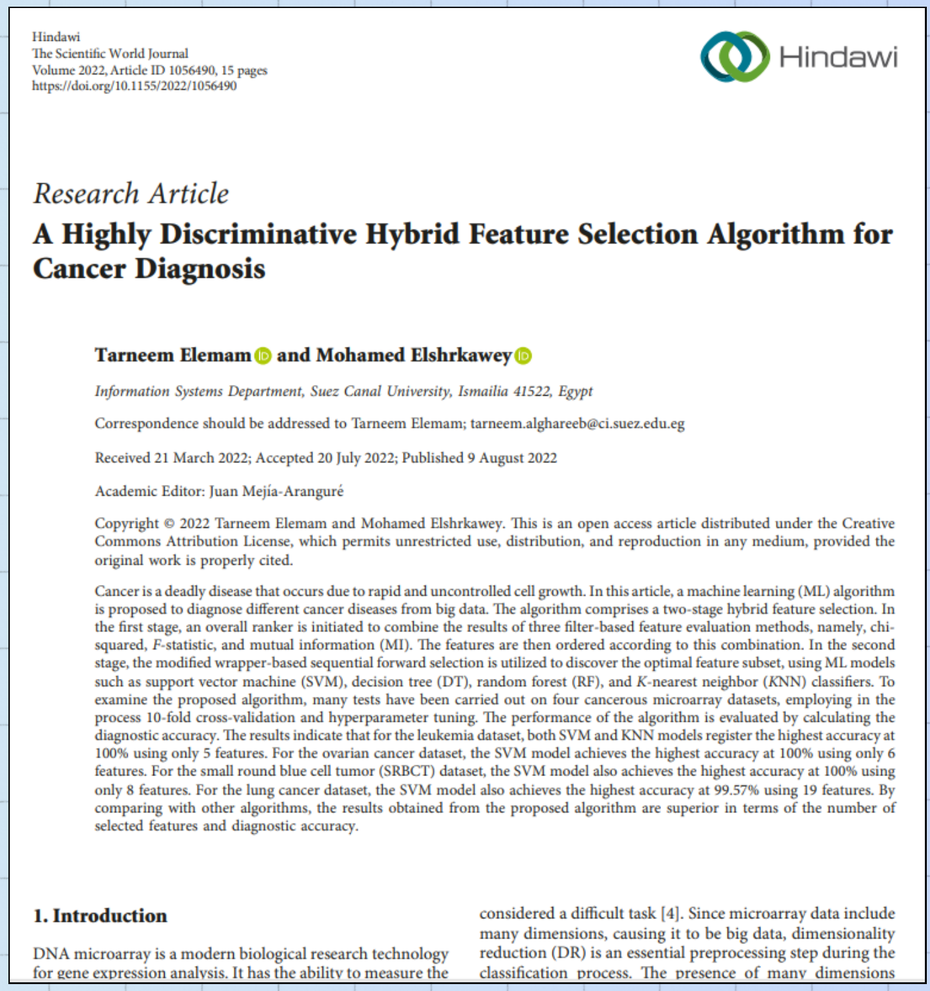
Ovarian Dataset Description

1. The ovarian cancer dataset used in this research was produced as a result of a study by Elemam, T., & Elshrkawey, M. (2022). A highly discriminative hybrid feature selection algorithm for cancer diagnosis. The Scientific World Journal, 2022, 1–15. https://doi.org/10.1155/2022/1056490
the paper accessible to researchers.

2. Gene selection has played an important role in cancer diagnosis and classification. 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 and high demensional data.

3. The mentioned dataset consists of 15155 genes (features), 253 observations samples and 2 classes.

4. The current observation group consisted of 162 people with the disease and 91 healthy people.



Hindawi
The Scientific World Journal
Volume 2022, Article ID 1056490, 15 pages
https://doi.org/10.1155/2022/1056490

Hindawi

Research Article

A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis

Tarneem Elemam and Mohamed Elshrkawey

Information Systems Department, Suez Canal University, Ismailia 41522, Egypt

Correspondence should be addressed to Tarneem Elemam; tarneem.alghareeb@ci.suez.edu.eg

Received 21 March 2022; Accepted 20 July 2022; Published 9 August 2022

Academic Editor: Juan Mejía-Aranguré

Cancer is a deadly disease that occurs due to rapid and uncontrolled cell growth. In this article, a machine learning (ML) algorithm is proposed to diagnose different cancer diseases from big data. The algorithm comprises a two-stage hybrid feature selection. In the first stage, an overall ranker is initiated to combine the results of three filter-based feature evaluation methods, namely, chi-squared, *F*-statistic, and mutual information (MI). The features are then ordered according to this combination. In the second stage, the modified wrapper-based sequential forward selection is utilized to discover the optimal feature subset, using ML models such as support vector machine (SVM), decision tree (DT), random forest (RF), and *K*-nearest neighbor (*K*NN) classifiers. To examine the proposed algorithm, many tests have been carried out on four cancerous microarray datasets, employing in the process 10-fold cross-validation and hyperparameter tuning. The performance of the algorithm is evaluated by calculating the diagnostic accuracy. The results indicate that for the leukemia dataset, both SVM and KNN models register the highest accuracy at 100% using only 5 features. For the ovarian cancer dataset, the SVM model achieves the highest accuracy at 100% using only 6 features. For the small round blue cell tumor (SRBCT) dataset, the SVM model also achieves the highest accuracy at 100% using only 8 features. For the lung cancer dataset, the SVM model also achieves the highest accuracy at 99.57% using 19 features. By comparing with other algorithms, the results obtained from the proposed algorithm are superior in terms of the number of selected features and diagnostic accuracy.

1. Introduction

DNA microarray is a modern biological research technology for gene expression analysis. It has the ability to measure the considered a difficult task [4]. Since microarray data include many dimensions, causing it to be big data, dimensionality reduction (DR) is an essential preprocessing step during the classification process. The presence of many dimensions

Ovarian Dataset

Current relation

Relation: Orarian — Attributes: 15155

Instances: 253 — Sum of weights: 253

Attributes

All | None | Inver

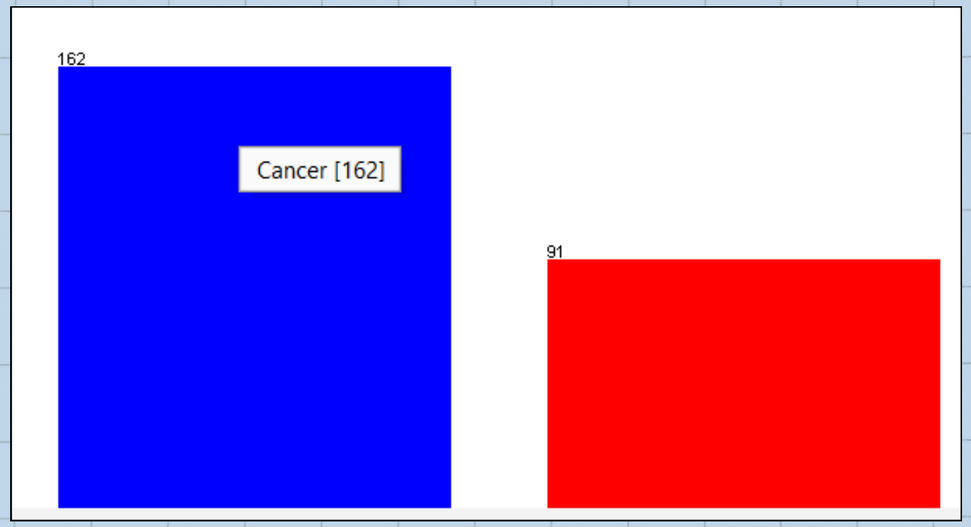
No.	Name
1	MZ-7.86E-05
2	MZ2.18E-07
3	MZ9.60E-05
4	MZ0.000366014
5	MZ0.000810195
6	MZ0.001428564
7	MZ0.002221123
8	MZ0.003187869
9	MZ0.004328805
10	MZ0.005643929
11	MZ0.007133241
12	MZ0.008796743
13	MZ0.010634432
14	MZ0.012646311
15	MZ0.014832378
16	MZ0.017192634
17	MZ0.019727078
18	MZ0.022435711
19	MZ0.025318532

Selected attribute

Name: Class — Type: Nominal

Missing: 0 (0%) — Distinct: 2 — Unique: 0 (0%)

No.	Label	Count	Weight
1	Cancer	162	162
2	Normal	91	91



The Ovarian dataset was downloaded as an Arff file in the Weka library, and Python was used to read and analyze these files. After reading the Arff file, the data was converted to Dataframe format in the Pandas library and the examination phase started. Then, the feature types were examined and it was found that only the features named 'Class' were categorical and the other features were numerical.

Column names are in the form of mass-to-charge ratios ('MZ'), specific column to numeric. Example: example assuming we want to convert the column with the name 'MZ0.000366014'.