

RECENT DEVELOPMENTS IN SOFT COMPUTING BASED TECHNIQUES FOR FEATURE SELECTION AND DISEASE CLASSIFICATION

Naiyar Iqbal* and Pradeep Kumar

Received: September 09, 2021; Revised: September 14, 2022; Accepted: September 27, 2022

Abstract

Computational prediction of diseases is vital in medical research that contributes to computer-aided diagnostics and helps doctors and medical practitioners in critical decision-making for various diseases such as bacterial and viral kinds of disease, including COVID-19 of the current pandemic situation. Feature selection techniques function as a preprocessing phase for classification and prediction algorithms. For disease prediction, these features may be the patient's clinical profiles or genomic features such as gene expression profiles from microarray and read counts from RNA-Seq. The performance of a classifier depends primarily on the selected features. In addition, genomic features are too large in numbers, resulting in the curse of dimensionality problem. In the last few years, several feature selection algorithms have been developed to overcome the existing problems to get rid of eliminating chronic diseases, such as various cancers, Zika virus, Ebola virus, and the COVID-19 pandemic. In this review article, we systematically associate soft computing-based approaches for feature selection and disease prediction by applying three data types: patients' clinical profiles, microarray gene expression profiles, and RNA-Seq sample profiles. According to related work, when the discussion took place, the percentage of medical data types highlighted through pictorial representation and the respective ratio of percentages mentioned were 52%, 27%, 9% and 12% for clinical symptoms, gene expression, MRI-Image and other data types such as signal or text-based utilized, respectively. We also highlight the significant challenges and future directions in this research domain.

Keywords: Clinical Symptom; Covid-19 Prediction; Disease Classification; Feature Selection; Gene Expression; Soft Computing

Introduction

We live in an era of rapidly changing global landscapes and local regional atmospheres. As a result, new viral and bacterial diseases, including cancer, have become serious threats to human life. Viral or bacterial diseases are contagious and spread from person to person or animal to person when a virus or bacterial microbe enters the body and

begins to multiply. In due course, the early prediction of diseases can protect human life by being aware of them to start proper diagnosis, care, and treatment. However, the early prediction of infectious diseases, such as dengue fever, Zika virus, Ebola virus, and cancers, is challenging (Khan *et al.*, 2017). Most developed viral and bacterial

Department of Computer Science and Information Technology, Maulana Azad National Urdu University (Central University), Hyderabad, Telangana, India. E-mail: naiyariqbal.rs@manuu.edu.in*; pradeep@manuu.edu.in

* Corresponding Author

disease prediction models are based on a patient's disease symptoms and clinical profiles. However, its prediction using - omic data (genomics, transcriptomic, proteomic, etc.) is still in its infancy stage. Biological -omic data, such as transcriptomic gene expression profiles of microarray as well as next-generation sequencing RNA-Seq, can be used to build for building a predictive model of various diseases. However, gene expression profiles face the curse of dimensionality issue. Therefore, a subset of reliable features must be collected before developing the prediction model.

Currently, the availability of prominent computational techniques for soft computing and tools for data analysis and prediction models require bioinformatics analysts and medical professionals to select the most convenient action plan to adapt to disease prediction issues. Soft computing offers various problem-solving techniques to handle data analysis and the development of prediction models (Bellazzi and Zupan, 2008; Liu and Liu, 2020).

The size of biological features, especially at the genetic level expressed in the form of gene structures, is in huge volume. Whole features are not so essential or relevant for every perspective. With the aim that the huge quantity of features conveys the issue of memory utilization, keeping the large dataset, it is difficult to choose the appropriate features for a particular problem. A feature selection model is an approach that includes the selection of a subset of appropriate attributes to enhance prediction model accuracy and eliminate redundant attributes (Inbarani *et al.*, 2014).

1.1 Significance and Motivation

The significance of feature selection activity can be seen in its application in the previous few years. For example, since the era of significant data started, feature selection procedures have been used to control the curse of dimensionality problems in the medical area for predicting various diseases.

According to PubMed publication repositories, the total number of publications with “*Feature Selection*” and “*Disease Prediction*” terms in the title or abstract from January 2000- December 2021 is 10867 (Figure 1). On the flip side, the aggregate number of publications with “*Soft Computing*” or “*Machine Learning*” terms in the title or abstract from January 2000 - December 2021 is 58938 (Figure 1). The previous decade justified that approximately a hundred-year contribution to PubMed is based on feature selection, disease prediction, soft computing, or machine learning. Using the literature search (Figure 2), we collected relevant work from prominent publication databases for this systematic research study.

1.2 Key Contributions

In this article, the key contributions are summarized as follows:

- The significance of topics covered based on year-wise publications related to feature selection and disease prediction methods was reported in PubMed from January 2000 to December 2021.
- An integrated framework for feature selection and disease prediction is proposed.
- Feature selection approaches with pros, cons, and techniques are presented along with a feasibility checklist for feature selection.
- The currently available benchmark toolkits for feature selection and disease prediction model development are also presented.
- Seven important challenges are discussed which lead to future research directions such as the curse of dimensionality problems, non-availability of negative class data, and cost of accessing a huge database.

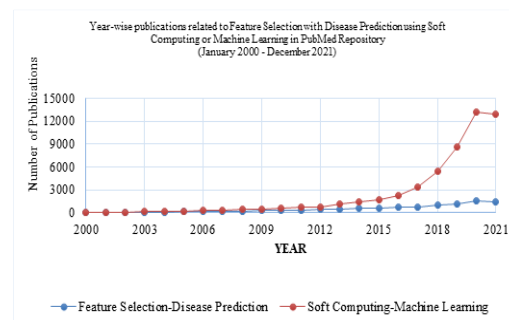


Figure 1. Year-wise publications related to feature selection and disease prediction methods reported in PubMed

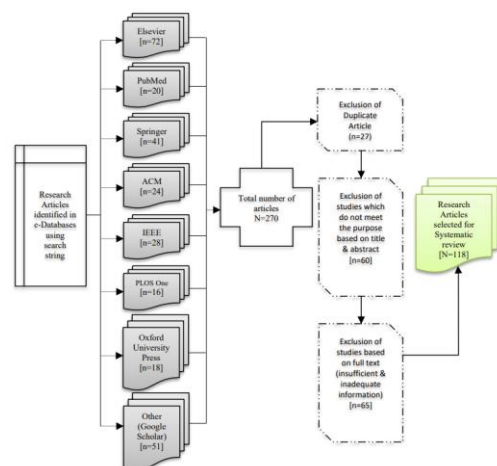


Figure 2. Literature search for the study

In this systematic review, the background of soft computing-based feature selection techniques for disease classification is explained in Section 1, along with the significance, motivation and key contributions of this work. Furthermore, various soft computing approaches ranging from simple to complex, as well as optimization techniques, are cited for readers. Related works are discussed in Section 2 and Section 3 proposes an integrated framework for feature selection and disease classification using experimental models. Section 4 is depicted five broad approaches to feature selection with their pros and cons with techniques applied. In Section 5, feasibility checklists for feature selection are presented. Finally, Section 6 is listed (Table 4) some prominent benchmark toolkits that are helpful for the implementation of the model. Furthermore, Section 7 is populated with 48 experimental works of various researchers that show model, data type, disease and accuracy. In Section 8, discussions focus on the implications and scope of the study. Section 9 points out seven critical challenges that further move toward future research directions. Finally, conclusions are summarized in Section 10.

Related Works

Soft computing is also well known as the *computational intelligence* (CI) technique through which computationally challenging tasks can be solved. Soft computing varies from traditional hard computing to tolerate imprecision, incomplete truth, uncertainty, and adjacency (Raza and Parveen, 2013). The primary techniques of soft computing are ranging from simple like Fuzzy logic (Lahsasna *et al.*, 2012), and probabilistic reasoning to complex machine learning techniques, namely Artificial Neural Networks (Abbass, 2002), Support Vector Machine (Jabeen *et al.*, 2018); Evolutionary computation (Xue *et al.*, 2016) like Genetic algorithm (Zhao *et al.*, 2011), Differential Evolution (De Falco, 2013), Genetic Programming (Hong and Cho, 2004), Gene Expression Programming (Zhang *et al.*, 2018), Cartesian Genetic Programming (Miller, 2019); Metaheuristic and Swarm Intelligence like Artificial Bee Colony (Hancer *et al.*, 2015), Ant Colony Optimization (Kabir *et al.*, 2012), Particle Swarm Optimization (Lin and Chen, 2009) and Artificial Fish Swarm Optimization (Wang *et al.*, 2014); Probabilistic and Statistical methods like Bayesian Network (Lin and Haug, 2008), Principal Component Analysis (Hariharan *et al.*, 2014) and Formal Concept Analysis (Raza, 2017); Hybrid methods like Neuro-Fuzzy (Azar and Hassanien, 2015), Neuro-Genetic (Rawat and

Burse, 2013), Neuro-Fuzzy-Genetic (Iqbal and Kumar, 2019), Artificial Neural Network- Ant Colony Optimization (Sivagaminathan and Ramakrishnan, 2007), Genetic Algorithm- Support Vector Machine (Avci, 2009).

The concerned researchers can perceive related works that survey the feature selection technique applied utmost in the area of research. Saey *et al.* (2007) explore a fundamental classification of traditional feature selection methods and explain the different applications of bioinformatics using feature selection methods. Thangavel and Pethalakshmi (2009) studied dimensionality reduction using rough set theory and their hybrid approach with the fuzzy set, neural network, and metaheuristic algorithms. Vergara and Estévez (2014) show a review on feature selection based on mutual information that describes a unifying theoretical framework and also discusses various open challenges in the area of feature selection. Finally, bolon-Canedo *et al.* (2014) present a review of applied feature selection techniques using microarray datasets and tabulated investigation outcomes aimed at cross-validation analysis with three classifiers, i.e., C4.5, Naïve Bayes, and SVM.

Proposed Integrated Framework

The primary focus of this study is to systematically review feature selection and disease prediction, which broadly comprises two phases: *The feature selection phase and the disease prediction phase*. In the subsequent section, we briefly define the tasks of each step in these phases. First, feature selection requires a massive database of high-dimensional features. Thereafter, the datasets are split into two partition training and testing datasets based on splitting criteria such as leave-one-out, m-fold, or any other (Figure 3).

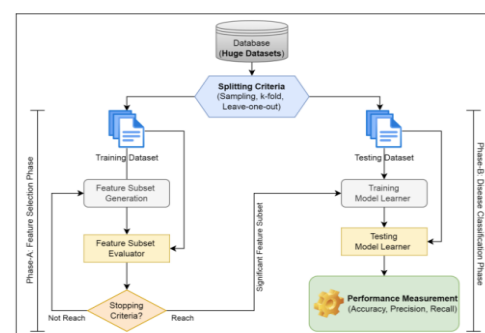


Figure 3. Integrated framework for the feature selection process and disease classification

The feature selection phase broadly consists of the following components: feature subset generation, feature subset evaluator, and stopping criteria. The prediction model phase broadly consists of the following components: training the model learner, testing the model learner, and performance measurement.

3.1 Feature Subset Generation

It is a heuristic-based search process, with every stage in the search space prescribing a candidate subset for estimation. Two main aspects influenced the perspective of the procedure.

Primary, one should choose the search initial point, which impacts the search fashion (i.e., forward, backward, bidirectional, random). In a forward manner, the search may begin with a blank set and progressively include features. In a backward fashion, it begins with a whole set and progressively eliminates features. In contrast, bidirectional learning begins in both directions by the inclusion and exclusion of features concurrently. Sometimes, the search may also begin randomly to select a subset to avoid being captured into local optima (Mehrpooya *et al.*, 2021). Additionally, a search approach should be chosen. Hence, various approaches have been analyzed, such as complete, sequential, and random searches.

3.2 Feature Subset Evaluation

A significant segment of any feature selection approach is the evaluation of the feature subsets. The evaluation functions evaluate how well

a particular feature subset can be in segregate among classes and can be classified into five broad feature selection strategies: filter, wrapper, embedded, ensemble, and hybrid.

3.3 Stopping Criteria

Feature selection strategies should be decided based on when to terminate searching into the feature subsets space. Few techniques prompt to predefine the number of selected features. On the flip side, some techniques follow the evaluation function strategy, such as whether to add or remove any feature which collects a better or an optimal feature subset. A stopping criterion regulates at what time the feature selection process should end. There are the following periods considered for stopping criteria such as i) Completion of search, ii) arrival of the specified threshold, iii) successive addition or elimination of any attribute that doesn't provide a better feature subset, and iv) an appropriate feature subset has been chosen.

3.4 Training Model Learner

In this phase, the best feature subset applies to the training process supplied by the feature selection phase. In machine learning, the training phase is primarily applied to train the model to produce correct outcomes.

3.5 Testing Model Learner

The trained model will be tested by applying trained datasets in this phase. Predictive Model

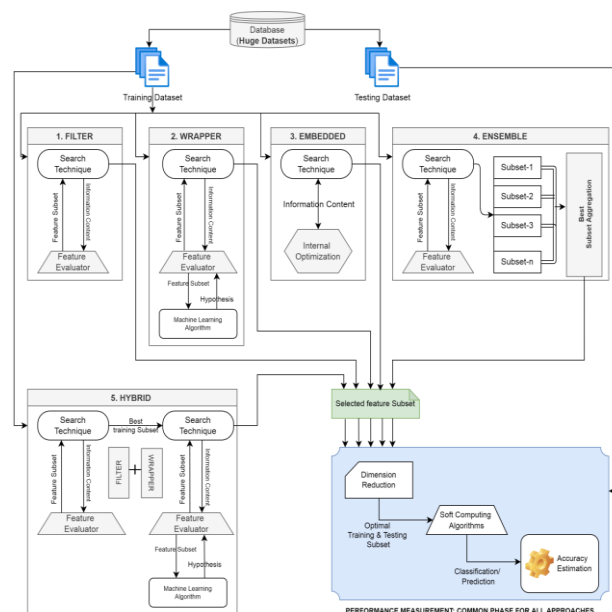


Figure 4. Broad types of feature selection approach

testing is a significant stage in the machine learning model. At this stage, the predictive model is tested spaciouly. It additionally incorporates the integration of the model.

3.6 Performance Measurement

After the training and testing phase, the model will be measured with the help of various parameters to check performance. The predictive model's performance was evaluated using various techniques and measurements using a machine learning-based confusion matrix such as accuracy, precision, recall and F1 measure etc.

Feature Selection Approach

Feature selection is a procedure for selecting a subset of attributes by using a group of substantial real attributes. The main aims of feature selection

are identifying relevant attributes, removing inappropriate attributes, and developing a better prediction system (Inbarani *et al.*, 2015). Using feature selection, two significant benefits can be achieved: a reduction in computational time and performance improvement of the model.

The goal of feature selection has three phases: performance improvement of prediction models, provide quicker and more economical predictive models, and provide superior comprehension of the fundamental procedure which produced the data.

In the subsequent section, we describe five feature selection approaches (Romalt and Kumar, 2020) that aim to identify the relevant attributes in various prominent algorithms. We also depict all five approaches in Figure 4. Table 1 presents the classification of feature selection approaches and their pros and cons. Table 2 lists some of the prominent feature selection algorithms.

Table 1. Taxonomy of feature selection approaches and their pros and cons

SEARCH APPROACH		PROS	CONS	EXAMPLE TECHNIQUES
FILTER	Univariate	- Well scalable with very high dimensional datasets - Fast and simple computational - Independent from a particular algorithm - Only once the feature selection process is done, then various classifiers can be estimated - Low time complexity	- Don't consider the collaboration with the classifier - Each feature is computed individually and so doesn't consider the feature dependencies - Performance reduction because of the absence of feature dependencies in contrast to another approach	- Relief algorithm - Focus algorithm - Peterson correlation coefficient - f-score - t-test - Chi-Square - Mutual Information - Euclidean Distance - Information Gain - Bayesian Networks
	Multivariate	- Incorporate feature dependency - Classifier independence - Better computational complexity	- Absence of classifier interaction - Less scalable and slower	- Correlation-based feature selection (CFS) - Markov Blanket Filter
WRAPPER	Deterministic	- Incorporate association amid model determination and searching attribute subset - Consider attribute dependence - Easy implementation and simple in a supervised learning approach - Execute comprehensive search - Producing optimum solutions	- Lower overfitting risk on inadequate features - Computation time on large features inclined to local optima - Classifier dependency - Computational exhaustive - Time complexity exponentially - No consideration of sufficient actions for redundancy elimination - Computational cost	- Sequential Forward Generation - Sequential Backward Generation - Sequential search - Suboptimal search method - Genetic Algorithm
	Randomized	- Incorporate classifier interaction - Presence of feature dependencies	- High overfitting risk on inadequate features - Lower inclined to local optima - Computationally intensive - Classifier dependency	- Metaheuristic search algorithm - Genetic algorithm - Simulated annealing - Randomized hill-climbing
EMBEDDED		- Connects with the classifier accomplishes computational complexity stronger in comparison to the wrapper approach - Feature dependency	- Classification dependency - Inclined to overfitting	- Decision tree - Weighted Naïve Bayes - SVM weight vector - Random forest
HYBRID		Integrates the qualities of several approaches	Time complexity may increase	A mixture of two or more algorithms
ENSEMBLE		- Lower inclined to overfitting - Higher flexibility and robustness upon high dimensional data	An ensemble of classifiers is hard to understand	- Bagging - Boosting - Random forest

Table 2. Some prominent feature selection algorithms

ALGORITHM	DESCRIPTION	REFERENCES
Relief	It selects a random sample from the feature set and decides its close hit and close miss relative to the feature set. Then, it follows the filter approach.	(Molina <i>et al.</i> , 2002) (Ang <i>et al.</i> , 2016)
Focus	It begins by estimating every individual attribute set, and after that, every group of two attributes, and so on. It breaks when an appropriately reliable solution is obtained. It also follows the filter approach.	(Molina <i>et al.</i> , 2002)
Las Vegas Filter (LVF)	LVF repetitively produces arbitrary feature subgroups and then calculates their assessment measurement. It was formally incorporated with the steadiness of the sample as an estimation measurement.	
Las Vegas Incremental (LVI)	LVI is founded because applying the entire sample to estimate the measure is essential.	
Sequential Floating Backward Selection (SFBS)	SFBS begins with whole feature sets. Then, after every step, it moves forward until the target function is reached.	
Branch and Bound (B&B)	B&B is an optimized search procedure in which a threshold value set by the operator stops the search at every node whose estimation is lesser than the threshold so that centrifugal branches are trimmed.	
Automatic Branch and Bound (ABB)	ABB is a kind of B&B in that a threshold value is assigned automatically.	
Quick Branch and Bound (QBB)	QBB is a hybridization of the LVF and ABB algorithms. The fundamental concept involves inventing a good beginning point for ABB with the help of the LVF. Then, it is predicted that ABB can effectively investigate the rest of the search space.	(Molina <i>et al.</i> , 2002) (Peng <i>et al.</i> , 2010)
Sequential Forward Selection (SFS)	SFS repeatedly include attributes to a starting subset to increase a specified measure, considering those attributes that previously existed in the solution. Besides, a well-organized list can also be acquired. It works well when the feature subset is smaller. It follows the wrapper approach.	
Sequential Backward Selection (SBS)	SBS is the backward corresponding algorithm of SFS. It works well when the feature subset is more significant. It also follows the filter approach.	(Molina <i>et al.</i> , 2002) (Peng <i>et al.</i> , 2010)
Sequential Floating Forward Selection (SFFS)	SFFS begins with empty feature sets. Then, after every step, it moves backwards till the target function is reached. In short, an attribute is initially included unconditional, then attributes are eliminated while the best feature subsets have been generated to their related size.	(Peng <i>et al.</i> , 2010)

4.1 Filter Approach

The filter approach is an unsupervised learning approach that selects the subset of features by looking at only the input with the essential information. In this approach, the classifier is not applied to attribute subset collection (Molina *et al.*, 2002). Instead, the preprocessing phase applies the typical characteristics of the training dataset to choose relevant attributes and eliminate inappropriate attributes (Peng *et al.*, 2010). Hence, filter approaches are free from the induction techniques that will utilize their outcome and can be joined with any such technique (Ang *et al.*, 2016). Univariate and multivariate are two classes of filter approaches. In univariate, every attribute is treated independently of the others that rank the feature by importance, and the user determines the ranking cutoff. The study of univariate attributes is the least difficult kind of investigation along the tracks since the attribute manages just a solitary amount that changes (Liu and Yu, 2005). It does not oversee causes or associations, and the primary inspiration in the back of the examination is to depict the information and find plans that occur inside it. In multivariate (Hira and Gillies, 2015), all features were considered simultaneously. This examination manages the investigation of multiple factors to comprehend the impact of factors on the reactions.

4.2 Wrapper Approach

The wrapper approach is a supervised learning style in which a selected subset of features is trained, and the estimation of error is done to validate the

dataset (Molina *et al.*, 2002). In this method, the search algorithm gives an attribute subset that is again used with a machine learning procedure, and prediction accuracy is obtained that is fed to the search algorithm, and after the whole module is completed repeatedly, a final feature subset that can use by the machine learning algorithm (Ang *et al.*, 2016). In this strategy, a classifier is applied as an estimation function.

The wrapper-based method can be deterministic and randomized. The same computational procedures are followed in deterministic algorithms to generate the same outcome on a specified input. The deterministic wrapper has less overfitting risk, is inclined to local optima, and has classifier dependency (Hira and Gillies, 2015).

Randomized algorithms take chances during the computation. This affects only the direction of the internal computing process. Pahikkala *et al.* (2012) developed greedy Regularized least-square (RLS) algorithm based on the wrapper approach for feature selection on a genome-wide level. Their experiment found a remarkably predicted subset of genetic variations in less amount period through wrapper and SVM techniques.

4.3 Embedded Approach

The embedded approach carries out feature selection during training time. First, it encompasses feature selection as fragments during training might be increasingly productive in various ways. Second, it utilizes the available data without dividing the

training data-obsessed by a training and validation dataset (Guyon and Elisseeff, 2003). Third, it arrives at a quicker result to avoid rehabilitating the prediction model from the start to each feature subset examined. Finally, the approach stimulates logical conjunctions to come up with an application of this embedded method. Strategies for promoting legitimate depictions give the articulate case of feature selection techniques embedded in a fundamental induction algorithm (Krishnapuram *et al.*, 2004). For these strategies, limited ordering explains the space of hypotheses, and algorithms typically apply this ordering to arrange their quest for definitive explanations (Liu *et al.*, 2020).

4.4 Ensemble Approach

The ensemble approach is a technique that expects to build a set of attribute subsets and afterwards generate an accumulated outcome from each subset. This approach depends on various sub-sampling procedures. A specific attribute selection process is executed on various

sub-samples and the acquired attributes converge to shape an increasingly steady subset (Abeel *et al.*, 2009). The presentation of attribute selection relies upon a solitary chosen subset. In this way, it is progressively adaptable and robust when managing the high dimensionality of data (Zomaya, 2013). Furthermore, this approach gives a superior estimation of the ideal subset or positioning of attributes by amassing the outcomes of various attribute selectors (Ye *et al.*, 2020).

4.5 Hybrid Approach

This approach is a hybrid of the filter and wrapper approaches that apply the advantages of both methods for selecting attribute subsets. It can be designed by merging two distinct strategies, the two approaches of a similar measure or two feature selection approaches. (Hsu *et al.*, 2011). It utilizes specific assessment measures in various inquiry stages to improve effectiveness and forecast execution with better computational performance.

Table 3. Feasibility checklist for feature selection

FEASIBILITY CHECKLIST	IF YES	IF NO
Domain knowledge	Create a better set of features.	Explore domain knowledge for building an effective model
Appropriate features	Applicable to use without or less normalization	Normalization needed
Features interdependence	Effectiveness of features correlation achieved	Associate features by expansion as much as possible
Pruning requires input features	Prune for reducing the computational cost, time of execution	Develop disjunctive features
Require features assess separately	Apply a feature ranking technique	Apply it somehow to find the axis of reference outcomes
Require predictor	Go	Stop
Suspect dirty data	find the outlier instances and remove	Higher accuracy by a model without outlier and noisy data
Knowledge of initial attempt	Apply non-linear predictor with feature subsets	Apply linear predictor
Novel model, computational resources	Perform a comparative analysis of various feature selection techniques with a novel model	Try to develop a novel model
Need stagnant solution	Re-sampling data subsets iteratively for high performance	Used only for specific data subset

Table 4. Benchmark Feature Selection Toolkit

TOOL NAME	ORGANISATION/ DEVELOPER	DEVELOPMENT LANGUAGE	REFERENCE
<i>General Purpose Feature Selection Simulators</i>			
WEKA	University of Waikato	Java	(Hall <i>et al.</i> , 2009)
FSLib	MathWorks	-	(Roffo, 2016)
Fast Correlation Based Filter	Arizona State University Lab	Java	(Senliol <i>et al.</i> , 2008)
sigFeature	Bioconductor	R	(Das <i>et al.</i> , 2018)
<i>Microarray Analysis Feature Selection Simulators</i>			
SAM (Significance Analysis of Microarrays)	Michael Seo	R, Excel	(Tusher <i>et al.</i> , 2001)
GALGO	Victor Trevino & Francesco Falciani	R	(Trevino and Falciani, 2006)
PCP (Pattern Classification Program)	Ljubomir Buturovic, MIT	C, C++	(Buturovic, 2006)
RankProd	Bioconductor	R	(Hong <i>et al.</i> , 2006)
Limma			(Smyth <i>et al.</i> , 2005)
Multtest			(Pollard <i>et al.</i> , 2005)
DGP			(Glez-Peña <i>et al.</i> , 2009)
ComparativeMarkerSelection (GenePattern)	University of California, MIT	Java, MATLAB, R	(Gould <i>et al.</i> , 2006)
GA/KNN	Leping Li	C	(Li <i>et al.</i> , 2004)
<i>Genomic Analysis Simulators</i>			
OmicsMarker	Bioconductor	R	(Determan and Determan, 2017)
Ropls			(Thévenot, 2016)

Feasibility Checklists

Some feasibility checklists or general guidelines listed in Table 3 could be followed while applying feature selection. The checklist covers all possible aspects which would be helpful to implement feature selectors followed by the development of machine learning models, especially with the availability of a huge dimension of feature space.

Benchmark Toolkits

The benchmark toolkits are applied to filter correlated features from a huge dataset. As a result, it handles the curse of dimensionality problems and enhances the accuracy of the prediction model. In Table 4, we listed some benchmark feature selection toolkits for general-purpose feature selection, microarray analysis feature selection, and genomic analyses developed by prominent organizations. Bioconductor found more participating organizations in the feature selector tool development in this listed toolkit. Furthermore, R language found a more applied programming language for the toolkit development.

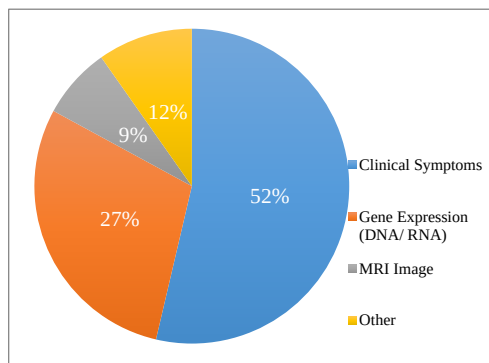


Figure 5. Distribution of contributed medical data types (in %)

Soft computing-based disease prediction models

We collected all related literature research works that applied various soft computing techniques for feature selection and disease prediction models from various prominent publications covering twenty-one years from 2000 to 2021. Table 5 shows the application of various disease predictions, along with feature selection techniques.

This article reviews soft computing-based techniques with feature selection for disease prediction and classification models. The performance of soft computing approaches was assessed based on their accuracy. The take-home conclusion of our review (Table 5) is that most cancerous diseases like colon, leukemia, breast cancer, lung cancer, and the novel coronavirus or COVID-19 are frequently applied ensemble approaches with feature selection algorithms for the development of predictive models which give the highest accuracy. Furthermore, primarily clinical symptoms-based datasets are used for the experiment. On the other hand, Gene expression-based data like DNA, RNA-Seq (Iqbal and Kumar, 2020), etc. are genetic level data that provide the highest level of prediction and classification of various diseases. The percentage distribution of experimental works performed by various researchers is also shown in Figure 5. 52% of collected experimental literature used clinical symptoms, whereas only 27% used gene expression datasets. Other studies based on MRI or other types of medical data used only 9% and 12%, respectively.

Our survey found that very few models have applied gene expression datasets to develop a prediction model. Therefore, further research and development are required in this direction. The curse of dimensionality is one of the significant challenges, especially at genetic level datasets addressed in the future direction.

Table 5. Related work in the direction of feature selection and disease prediction

MODEL	DATA TYPE	DISEASE	ACCURACY (%)	REFERENCE	
kNN-SFFS GA-kNN	Clinical symptoms	▪ Thyroid Dysfunction	97.99	(Raymer <i>et al.</i> , 2000)	
		▪ Acute Appendicitis	98.48		
CLIFF	Gene expression	▪ Type-I Leukemias (ALL)	90.30	(Xing and Karp, 2001)	
		▪ Type-II Leukemias (AML)	38.70		
IB1		▪ Colon	91.94	(Inza <i>et al.</i> , 2002)	
		▪ Leukemia	100.00		
		▪ NCI-60	85.00		
Ensemble		▪ Leukemia	97.10	(Cho and Won, 2003)	
		▪ Colon	93.60		
		▪ Lymphoma	96.00		
ENN		ALL-AML Leukemia		100.00	(Liu <i>et al.</i> , 2004a)
		▪ Lung Cancer		100.00	
		▪ Prostate Cancer		97.00	
		▪ DLBCL		98.00	
		▪ Ovarian Cancer		98.82	
		▪ Colon Tumor		91.94	
		▪ MLL Leukemia			

Table 5. Related work in the direction of feature selection and disease prediction

MODEL	DATA TYPE	DISEASE	ACCURACY (%)	REFERENCE
GMDH-NN	Clinical symptoms	- Breast Cancer (Wisconsin) - Heart disease (Cleveland)	97.50 85.00	(Abdel-Aal, 2005)
FS-SFS		- BUPA Live - BCW - Heart Disease - PI Diabetes - PO Patients - WDBC	70.20 96.30 78.30 74.50 71.80 93.00	(Liu and Zheng, 2006)
FW-AIRS		- Heart disease - Hepatitis disease	92.59 81.82	(Polat and Güneş, 2007)
RBFS-RBF NN		- ALL-AML Leukemia - Lung Cancer - Prostate Cancer	99.89 98.84 99.62	(Chiang and Ho, 2008)
AR-NN	Clinical symptoms	Erythematous-squamous diseases	98.61	(Karabatak and Ince, 2009)
MOGA	Gene expression	- Leukemia - Colon - Lymphoma	98.03 89.58 96.05	(García-Nieto <i>et al.</i> , 2009)
NB	MRI Image	Brain Atrophy	92.00	(Plant <i>et al.</i> , 2010)
SVM-RBF	3D SPECT (single-photon emission computed tomography) brain Image	Alzheimer	96.91	(Illán <i>et al.</i> , 2011)
PCA-LSSVM	Clinical Symptoms	Hepatitis	96.12	(Çalışır and Dogantekin, 2011)
RF-IBk		Parkinson Disease	97.00	(Ozcift, 2012)
RBF		Breast Cancer	95.85	
MLP		Diabetes	75.39	(Chen <i>et al.</i> , 2012)
PSO-SVM		Breast Cancer	99.30	
AR-PCA/PLS-SVM	- Single-photon emission computed tomography (SPECT) - Positron emission tomography (PET)	Alzheimer's Disease	91.75 90.00	(Chaves <i>et al.</i> , 2012)
mRMR-PSO-WSVM	Gene expression	- Leukemia - Colon	100.00 93.55	(Abdi <i>et al.</i> , 2012)
GA-BN	Clinical symptoms	Erythematous-squamous diseases	99.20	(Özçift and Gülsen, 2013)
RS-ELM		Hepatitis Disease	100.00	(Kaya and Uyar, 2013)
MABC-FC		Pima Indian Diabetes	84.21	(Beloufa and Chikh, 2013)
ANFCLH	Clinical and Histopathological symptoms	Erythematous-squamous diseases	95.77	(Azar <i>et al.</i> , 2013)
FMM-CART-RF	Clinical symptoms	- Breast Cancer - Pima Indian Diabetes - Liver Disorders	98.84 78.39 95.01	(Seera and Lim, 2014)
CS-PSO-SVM		- Heart Disease - Breast Cancer	85.00 91.33	(Liu and Fu, 2014)
RS-BPNN		- Hepatitis - Breast Cancer - Heart Disease	97.30 98.60 90.40	(Nahato <i>et al.</i> , 2015)
BN		Prostate cancer	94.11	(Raza and Hasan, 2015)
FL-DT	Clinical symptoms	Coronary Heart Disease	69.51	(Kim <i>et al.</i> , 2015)
GA-ANFIS		Dermatological Diseases	98.00	(Fazlic <i>et al.</i> , 2015)
DW-S2ML	MRI Image	Alzheimer's Disease	80.11	(Suk <i>et al.</i> , 2016)
IPCA-EM-ISVM	Clinical symptoms	Diabetes	97.95	(Nilashi <i>et al.</i> , 2016)
LLE-NRS	Gene expression	Acute leukemia	94.74	(Sun <i>et al.</i> , 2016)
ASCI RS-GA-NN	Clinical Symptoms	- Breast Cancer - Fertility - Heart Disease	95.33 97.61 93.04	(Sudha, 2017)
dgSeq	Gene expression	- Breast cancer - Thyroid cancer - Alzheimer's disease	88.00 83.00 80.00	(Luo <i>et al.</i> , 2017)
EWNN	Mammographic images, Biomedical voice measurements, Speech recordings	- Digital Database for Screening Mammography - Little's Parkinson's Disease - Sakar's Parkinson's Disease	95.50 100.00 90.00	(Khan <i>et al.</i> , 2018)
SDPSO-SVM-PCA	MRI Image	Alzheimer's Disease	76.86	(Zeng <i>et al.</i> , 2018)
OCFA	Speech PD VoicePD HandPD Meander HnadPD Spiral	Parkinson's disease	92.19 83.48 87.12 88.46	(Gupta <i>et al.</i> , 2018)
EBO-HSIC-SVM	Gene expression	Cervical cancer	90.00	(Geeitha <i>et al.</i> , 2018)
LogitBoost	Clinical symptoms	Dengue Outbreak	92.00	(Iqbal and Islam, 2019)
LR-MLP-NB-MVR		Heart Disease	88.88	(Raza, 2019)
PCA-Relief-Adaptive SVM		Chronic Kidney Disease	97.48	(Divya and Singh, 2019)

Table 5. Related work in the direction of feature selection and disease prediction

MODEL	DATA TYPE	DISEASE	ACCURACY (%)	REFERENCE
multi-kernel SVM		▪ Thyroid Disease	97.49	(Shankar <i>et al.</i> , 2020)
HFSM-EKNN		▪ COVID-19	96.00	(Shaban <i>et al.</i> , 2020)
AlexNet/ CNN		▪ COVID-19	79.00	(El-Kenawy <i>et al.</i> , 2020)
HLBDA		▪ COVID-19	92.21	(Too and Mirjalili, 2021)
ARFS		▪ Diabetes	73.83	(Huang <i>et al.</i> , 2020)
LAGOA		▪ Statlog (Heart)	88.24	(Dey <i>et al.</i> , 2021)
		▪ SPECTF Heart	88.21	
		▪ Breast Cancer (Wisconsin)	99.26	
		▪ Breast Cancer (Diagnosis)	98.46	
		▪ Lung Cancer	85.71	
		▪ Hepatitis	93.00	
SVM-RBF	MRI Image	▪ Alzheimer's Disease	96.82	(Divya and Kumari, 2021)
CSE-PCA Hoeffding Tree	Gene expression - RNA-Seq	▪ COVID-19	99.21	(Iqbal and Kumar, 2021)

Discussion

Innovative feature selection will be seen where productive data devaluation has to be handled when our capability to occupy and keep data far away from our capacity for processing and applying. Feature selection can support us in the directing a suitable part of data and improving our capability to process the data (Ahmad *et al.*, 2018). As feature selection is a data preprocessing stage, changes should be made for classical techniques that need various database scanning and random access to data (Bolón-Canedo *et al.*, 2015).

Feature selection is a vibrant area firmly associated with data mining, other data processing techniques, and soft computing techniques. This article endeavours to study this quick-growing field of research, demonstrate specific emphatic applications and pay attention to fascinating trends and challenges.

There is future scope to identify the biological significance of feature subsets after applying soft computing techniques instead of only considering the model's performance.

The performance requires to be assessed not only based on classification accuracy but also to assess the metrics such as sensitivity and specificity. Additionally, we have identified in our study that some of the important challenging points lead to the future research direction. Out of all challenges and future directions mentioned in the subsequent section, the dimension reduction problem is one of the crucial issues to feature selection from the massive feature vector in the field of genetic data analysis.

Challenges and Future Scope

We identified seven significant challenges based on the systematic literature survey of various studies, leading to future research directions. The *curse of*

dimensionality is one of the biggest challenges, especially in disease prediction at the genetic level.

9.1 Curse of Dimensionality

The curse of dimensionality is mentioned as a non-instinctive feature of data captured when handling higher dimensional space, which is significantly associated with the utilization and explanation of distance and volume. It fundamentally means that the error grows with the rise in the number of features (Pavlenko, 2003). Furthermore, it implies that models are challenging to design in high dimensions and often have an exponential time execution in higher dimensions. Conceptually, it provides more information to be stored in a high dimension of features, whereas practically, it hardly supports it because of noise and redundant data.

9.2 Non-availability of negative class data

Gathering negative training samples is particularly fragile and difficult because negative training samples should consistently describe the universal dataset for the exclusion of case (+ive) class and manually gathered control (-ive) training samples possibly will be biased because of a person's accidental prejudice, which might be harmful to classification accuracy (Iqbal and Kumar, 2022). The issue is that the non-accessibility of negative class information may prompt a bias assessment of any predictive model. However, perception normally considers the non-accessibility of negative class datasets.

9.3 Data imbalance issue

Data imbalance is a general issue in classification where there is an inappropriate proportion of observations in every class. Data imbalance can be found in a wide range of regions comprises clinical diagnosis, spam filter, and fraud detection. Imbalance data generally mention a classification issue where the distribution of

observations per class is unequal (Yin *et al.*, 2013). There are two main aspects; the majority class has a massive volume of data or observations for one class; whereas the minority class has less data or observations for another class.

9.4 Selection of active feature

Active feature selection stimulates the plan to choose samples for feature selection actively. It prevented a purely random sample approach and utilized a selective sample approach that uses the benefit of data properties when choosing samples. The main logic of the selective sample approach is to choose only those samples with the highest possibilities to help consider the feature significance (Liu *et al.*, 2004b). A few selective sampling techniques target better or equally good feature selection outcomes with fewer samples.

9.5 Cost of accessing huge database

The cost mentions the costs of gaining instances for individual features rather than the cost of error types, as in cost-sensitive learning or the cost of obtaining labels for samples, such as active learning (Florescu and Kossmann, 2009). Various feature selection techniques do not include exclusive feature costs. However, such cost measurements can constitute the correct cost of feature groups.

9.6 Scalability of dataset and Model

The scalability of most existing feature selection techniques might be a risk with the enormous increase in dataset volume. In numerous technical, commercial, and medical applications, data are typically scaled in huge volumes, such as in terabytes or more (Li and Liu, 2017). Generally, datasets in the size of a terabyte or more are difficult to store in memory directly and consequently bounds the serviceability of almost all attribute selection techniques. As a result, its effectiveness declines rapidly or becomes computationally infeasible.

9.7 Stability of algorithm

The stability of the algorithms is also significant attention when emerging novel feature selection techniques. However, exploring stability for unsupervised feature selection techniques is considerably more troublesome compared to supervised techniques (Li and Liu, 2017). The reason is the lack of sufficient earlier information about the cluster assembly of the data in the unsupervised feature selection. Therefore, it is undefined if the new data sample generates, belongs to any present clusters or requires new clusters.

Conclusion

In this systematic review, we described various aspects of feature selection and disease prediction using soft computing techniques to solve complex tasks. With the help of soft computing and its core strategies which vary from normal to complex techniques such as fuzzy logic, naïve based, Artificial Neural Networks, Support Vector Machine, etc., and a few evolutionary techniques like genetic algorithm, Differential Evolution, and gene expression programming also came into light. Furthermore, various optimisation techniques have been developed to perform the tasks at an optimum level, such as Ant Colony Optimization, Artificial Bee Colony, and Particle Swarm Optimization.

A general guideline or feasibility for feature selection is mentioned in the study. Besides that, we have listed various benchmark toolkits and datasets utilized for feature selection and disease prediction. We have also pointed out various emerging challenges and future direction that needs to resolve the problems in the near future to develop better prediction models.

Based on the review of related work, we pointed out some gaps that need to be addressed in the near future. For example, many researchers have proposed different feature selection and disease prediction models using soft computing techniques.

Gap (1) Most models are based on the clinical dataset, and a few literatures reported the application of microarray and RNA-Seq based on the prediction.

Gap (2) Most effective working models are “black-box” in nature. Therefore, we must also infer the prediction rules from these black-box models. Hence, there is a need to develop an accurate method for feature selection and disease prediction using microarray and RNA-Seq based gene expression data.

Gap (3) There are seven challenges discussed above that need to be resolved, such as the *curse of dimensionality* problem is one of the crucial issues at the genetic level.

With the prospects of resolving these issues, we can develop a high-throughput prediction model to combat existing diseases for prognosis and diagnosis.

Abbreviations

ANFCLH	: Adaptive Neural Fuzzy Classifier with Linguistic Hedges
ANFIS	: Artificial Neural Fuzzy Inference System

AR	:	Association Rules	NN	:	Neural Network
ARFS	:	Association Rule Feature Selection	OCFA	:	Optimised Cuttlefish Algorithm
ASCI	:	Application-Specific Intelligent Computing	PCA	:	Principal Component Analysis
BN	:	Bayesian Network	PLS	:	Partial Least Square
CLIFF	:	Clustering via Iterative Feature Filtering	PSO	:	Particle Swam Optimisation
CNN	:	Convolutional Neural Network	RBFS- RBF	:	Rough-Based Feature Selection- Radial Basis Function
CS	:	Cuckoo Search	RF-IBk	:	Rotation Forest-Instance Based Learner
CSE	:	Classifier Subset Evaluator	RS-BNN	:	Rough Sets-Back propagation Neural Network
dgSeq	:	Disease Gene Prediction Algorithm	RS-ELM	:	Rough Set- Extreme Learning Machine
DW-S2ML	:	Deep Weighted-Sparse Multi-Task Learning	SDPSO	:	Switching Delayed- PSO
EBO-HSIC	:	Enhanced Bat Optimization Algorithm with Hilbert-Schmidt Independence Criterion	SFFS	:	Sequential Floating Forward Selection
ENN	:	Ensemble Neural Network	SVM	:	Support Vector Machine
EWNN	:	Evolutionary Wavelet Neural Networks	WSVM	:	Weighted SVM
FL-DT	:	Fuzzy Logic- Decision Tree			
FMM-CART-RF	:	Fuzzy Min-Max Neural Network-Classification and Regression Tree-Random Forest			
FS-SFS	:	Filtered and Supported Sequential Forward Search			
FW-AIRS	:	Fuzzy Weighted- Artificial Immune Recognition System			
GA	:	Genetic Algorithm			
GMDH	:	Group Method of Data Handling			
HFSM-EKNN	:	Hybrid Feature Selection Methodology- Enhanced kNN			
HLBDA	:	Hyper Learning Binary Dragonfly Algorithm			
IPCA-EM-ISVM	:	Incremental PCA- Expectation-Maximization- Incremental SVM			
kNN	:	k- Nearest Neighbor			
LAGOA	:	Learning Automata based Grasshopper Optimization Algorithm			
LLE-NRS	:	Locally Linear Embedding and Neighborhood Rough Set			
LR-MLP-NB-MVR	:	Logistic Regression, Multilayer Perceptron, Naive Bayes, Majority Voting Rule			
LSSVM	:	Least Square SVM			
MABC-FC	:	Modified Artificial Bee Colony- Fuzzy Classifier			
MLP	:	Multi-Layer Perceptron			
MOGA	:	Multi Objective Genetic Algorithm			
mRMR	:	Minimum Redundancy Maximum Relevance			
NB	:	Naïve Bayes			

References

- Abbass, H.A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine.*, 25(3):265-281. [https://doi.org/10.1016/S0933-3657\(02\)00028-3](https://doi.org/10.1016/S0933-3657(02)00028-3)
- Abdel-Aal, R.E. (2005). GMDH-based feature ranking and selection for improved classification of medical data. *Journal of Biomedical Informatics.*, 38(6):456-468. <https://doi.org/10.1016/j.jbi.2005.03.003>
- Abdi, M.J., Hosseini, S.M., and Rezghi, M. (2012). A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification. *Computational and mathematical methods in medicine*, 2012. <https://doi.org/10.1155/2012/320698>
- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics.*, 26(3):392-398. <https://doi.org/10.1093/bioinformatics/btp630>
- Ahmad, W.M.T.W., Ab Ghani, N.L., and Drus, S.M. (2018). Data mining techniques for disease risk prediction model: A systematic literature review. In *International Conference of Reliable Information and Communication Technology*. Springer, Cham. p. 40-46. https://doi.org/10.1007/978-3-319-99007-1_4
- Ang, J.C., Mirzal, A., Haron, H., and Hamed, H.N.A. (2016). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics.*, 13(5):971-989. <https://doi.org/10.1109/TCBB.2015.2478454>
- Avci, E. (2009). A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. *Expert Systems with Applications.*, 36(7):10,618-10,626. <https://doi.org/10.1016/j.eswa.2009.02.053>
- Azar, A.T., and Hassanien, A.E. (2015). Dimensionality reduction of medical big data using neural-fuzzy classifier. *Soft computing.*, 19(4):1,115-1,127. <https://doi.org/10.1007/s00500-014-1327-4>
- Azar, A.T., El-Said, S.A., Balas, V.E., and Olariu, T. (2013). Linguistic hedges fuzzy feature selection for differential diagnosis of Erythematous-Squamous diseases. In *Soft computing applications* (pp. 487-500). Springer, Berlin,

- Heidelberg. https://doi.org/10.1007/978-3-642-33941-7_43
- Bellazzi, R., and Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics.*, 77(2):81-97. <https://doi.org/10.1016/j.ijmedinf.2006.11.006>
- Beloufa, F., and Chikh, M.A. (2013). Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. *Computer methods and programs in biomedicine.*, 112(1):92-103. <https://doi.org/10.1016/j.cmpb.2013.07.009>
- Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J.M., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences.*, 282:111-135. <https://doi.org/10.1016/j.ins.2014.05.042>
- Bolón-Canedo, V., Sánchez-Marono, N., and Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems.*, 86:33-45. <https://doi.org/10.1016/j.knsys.2015.05.014>
- Buturovic, L.J. (2006). PCP-Pattern Classification Program, version 2.2 User's Guide.
- Çalişır, D., and Dogantekin, E. (2011). A new intelligent hepatitis diagnosis system: PCA-LSSVM. *Expert Systems with Applications.*, 38(8):10,705-10,708. <https://doi.org/10.1016/j.eswa.2011.01.014>
- Chaves, R., Ramírez, J., Górriz, J.M., Puntonet, C.G., and Alzheimer's Disease Neuroimaging Initiative. (2012). Association rule-based feature selection method for Alzheimer's disease diagnosis. *Expert Systems with Applications.*, 39(14):11,766-11,774. <https://doi.org/10.1016/j.eswa.2012.04.075>
- Chen, H.L., Yang, B., Wang, G., Wang, S.J., Liu, J., and Liu, D. Y. (2012). Support vector machine based diagnostic system for breast cancer using swarm intelligence. *Journal of medical systems.*, 36(4):2,505-2,519. <https://doi.org/10.1007/s10916-011-9723-0>
- Chiang, J.H., and Ho, S.H. (2008). A combination of rough-based feature selection and RBF neural network for classification using gene expression data. *IEEE transactions on nanobioscience.*, 7(1):91-99. <https://doi.org/10.1109/TNB.2008.2000142>
- Cho, S.B., and Won, H.H. (2003). Machine learning in DNA microarray analysis for cancer classification. In *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics.*, 19:189-198
- Das, P., Roychudhury, S., and Tripathy, S. (2018). sigFeature: Significant feature selection using SVM-RFE and t-statistic. <https://doi.org/10.24870/cjb.2017-a22>
- De Falco, I. (2013). Differential Evolution for automatic rule extraction from medical databases. *Applied soft computing.*, 13(2):1,265-1,283. <https://doi.org/10.1016/j.asoc.2012.10.022>
- Determan Jr, C.E., and Determan Jr, M.C.E. (2017). Package 'OmicsMarker'.
- Dey, C., Bose, R., Ghosh, K.K., Malakar, S., and Sarkar, R. (2021). LAGOA: Learning automata-based grasshopper optimisation algorithm for feature selection in disease datasets. *Journal of Ambient Intelligence and Humanized Computing.*, p. 1-20. <https://doi.org/10.1007/s12652-021-03155-3>
- Divya, R., and Kumari, R.S.S. (2021). Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. *Neural Computing and Applications.*, p. 1-10. <https://doi.org/10.1007/s00521-020-05596-x>
- El-Kenawy, E.S.M., Ibrahim, A., Mirjalili, S., Eid, M.M., and Hussein, S.E. (2020). Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. *IEEE Access.*, 8, 179317-179335. <https://doi.org/10.1109/ACCESS.2020.3028012>
- Fazlic, L.B., Avdagic, K., and Omanovic, S. (2015). GA-ANFIS Expert System Prototype for Prediction of Dermatological Diseases. In *MIE* (pp. 622-626).
- Florescu, D., and Kossmann, D. (2009). Rethinking cost and performance of database systems. *ACM Sigmod Record.*, 38(1):43-48. <https://doi.org/10.1145/1558334.1558339>
- García-Nieto, J., Alba, E., Jourdan, L., and Talbi, E. (2009). Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. *Information Processing Letters.*, 109(16):887-896. <https://doi.org/10.1016/j.ipl.2009.03.029>
- Geeitha, S., and Thangamani, M. (2018). Incorporating EBO-HSIC with SVM for Gene Selection Associated with Cervical Cancer Classification. *Journal of medical systems.*, 42(11):225. <https://doi.org/10.1007/s10916-018-1092-5>
- Glez-Peña, D., Álvarez, R., Díaz, F., and Fdez-Riverola, F. (2009). DFP: A Bioconductor package for fuzzy profile identification and gene reduction of microarray data. *BMC Bioinformatics.*, 10(1):1-8. <https://doi.org/10.1186/1471-2105-10-37>
- Gould, J., Getz, G., Monti, S., Reich, M., and Mesirov, J.P. (2006). Comparative gene marker selection suite. *Bioinformatics.*, 22(15):1,924-1,925. <https://doi.org/10.1093/bioinformatics/btl196>
- Gupta, D., Julka, A., Jain, S., Aggarwal, T., Khanna, A., Arunkumar, N., and de Albuquerque, V.H.C. (2018). Optimised Cuttlefish Algorithm for Diagnosis of Parkinson's Disease. *Cognitive Systems Research.* <https://doi.org/10.1016/j.cogsys.2018.06.006>
- Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research.*, 3(Mar), 1,157-1,182.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter.*, 11(1):10-18. <https://doi.org/10.1145/1656274.1656278>
- Hancer, E., Xue, B., Karaboga, D., and Zhang, M. (2015). A binary ABC algorithm based on advanced similarity scheme for feature selection. *Applied Soft Computing.*, 36:334-348. <https://doi.org/10.1016/j.asoc.2015.07.023>
- Hariharan, M., Polat, K., and Sindhu, R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. *Computer methods and programs in biomedicine.*, 113(3):904-913. <https://doi.org/10.1016/j.cmpb.2014.01.004>
- Hira, Z.M., and Gillies, D.F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics.*, 2015. <https://doi.org/10.1155/2015/198363>
- Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L., and Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics.*, 22(22):2,825-2,827. <https://doi.org/10.1093/bioinformatics/btl476>
- Hong, J.H., and Cho, S.B. (2004). Lymphoma cancer classification using genetic programming with SNR features. In *European Conference on Genetic Programming.* Springer, Berlin, Heidelberg. P. 78-88. https://doi.org/10.1007/978-3-540-24650-3_8
- Hsu, H.H., Hsieh, C.W., and Lu, M.D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications.*, 38(7):8,144-8,150. <https://doi.org/10.1016/j.eswa.2010.12.156>

- Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., ... and Li, J. (2020). Sample imbalance disease classification model based on association rule feature selection. *Pattern Recognition Letters.*, 133:280-286. <https://doi.org/10.1016/j.patrec.2020.03.016>
- Illán, I.A., Górriz, J.M., López, M.M., Ramírez, J., Salas-Gonzalez, D., Segovia, F., and Puntónet, C.G. (2011). Computer-aided diagnosis of Alzheimer's disease using component-based SVM. *Applied Soft Computing.*, 11(2):2,376-2,382. <https://doi.org/10.1016/j.asoc.2010.08.019>
- Inbarani, H.H., Azar, A.T., and Jothi, G. (2014). Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Computer methods and programs in biomedicine.*, 113(1):175-185. <https://doi.org/10.1016/j.cmpb.2013.10.007>
- Inbarani, H.H., Bagyamathi, M., and Azar, A.T. (2015). A novel hybrid feature selection method based on rough set and improved harmony search. *Neural Computing and Applications.*, 26(8):1,859-1,880. <https://doi.org/10.1007/s00521-015-1840-0>
- Inza, I., Sierra, B., Blanco, R., and Larrañaga, P. (2002). Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent and Fuzzy Systems.*, 12(1):25-33.
- Iqbal, N. and Kumar, P. (2019). I-NFG: An integrated neuro-fuzzy-genetic based soft computing techniques for feature selection and disease prediction using gene expression. *Journal of Applied Computing.*, 4(1):1-8.
- Iqbal, N., and Islam, M. (2019). Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. *Informatica.*, 43(3). <https://doi.org/10.31449/inf.v43i3.1548>
- Iqbal, N., and Kumar, P. (2020). A framework for the RNA-Seq based classification and prediction of disease. In *ICDSMLA 2019* (pp. 74-81). Springer, Singapore. https://doi.org/10.1007/978-981-15-1420-3_8
- Iqbal, N., and Kumar, P. (2021, November). Coronavirus Disease Predictor: An RNA-Seq based pipeline for dimension reduction and prediction of COVID-19. In *Journal of Physics: Conference Series.*, 2,089(1):012025. IOP Publishing. <https://doi.org/10.1088/1742-6596/2089/1/012025>
- Iqbal, N., and Kumar, P. (2022). Integrated COVID-19 predictor: Differential expression analysis to reveal potential biomarkers and prediction of coronavirus using RNA-Seq profile data. *Computers in Biology and Medicine.*, 147:105684. <https://doi.org/10.1016/j.combiomed.2022.105684>
- Jabeen, A., Ahmad, N. and Raza, K. (2018). Machine Learning-based State-of-the-art Methods for the Classification of RNA-Seq Data. In: Dey N., Ashour A., Borra S. (eds) *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics*, Springer., 26:133-172. https://doi.org/10.1007/978-3-319-65981-7_6
- Kabir, M.M., Shahjahan, M., and Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. *Expert Systems with Applications.*, 39(3):3,747-3,763. <https://doi.org/10.1016/j.eswa.2011.09.073>
- Karabatak, M., and Ince, M.C. (2009). A new feature selection method based on association rules for diagnosis of erythematous-squamous diseases. *Expert Systems with Applications.*, 36(10):12,500-12,505. <https://doi.org/10.1016/j.eswa.2009.04.073>
- Kaya, Y., and Uyar, M. (2013). A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. *Applied Soft Computing.*, 13(8):3,429-3,438. <https://doi.org/10.1016/j.asoc.2013.03.008>
- Khan, F.N., Qazi, S., Tanveer, K., and Raza, K. (2017). A review on the antagonist Ebola: A prophylactic approach. *Biomedicine and Pharmacotherapy.*, 96:1,513-1,526. <https://doi.org/10.1016/j.biopha.2017.11.103>
- Khan, M.M., Mendes, A., and Chalup, S.K. (2018). Evolutionary Wavelet Neural Network ensembles for breast cancer and Parkinson's disease prediction. *PloS one.*, 13(2):e0192192. <https://doi.org/10.1371/journal.pone.0192192>
- Kim, J., Lee, J., and Lee, Y. (2015). Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. *Healthcare informatics research.*, 21(3):167-174. <https://doi.org/10.4258/hir.2015.21.3.167>
- Krishnapuram, B., Carin, L., and Hartemink, A. (2004). 1 Gene expression analysis: Joint feature selection and classifier design. *Kernel Methods in Computational Biology.*, p. 299-317.
- Lahsasna, A., Ainon, R.N., Zainuddin, R., and Bulgiba, A. (2012). Design of a fuzzy-based decision support system for coronary heart disease diagnosis. *Journal of medical systems.*, 36(5):3,293-3,306. <https://doi.org/10.1007/s10916-012-9821-7>
- Li, J., and Liu, H. (2017). Challenges of feature selection for big data analytics. *IEEE Intelligent Systems.*, 32(2):9-15. <https://doi.org/10.1109/MIS.2017.38>
- Li, L., Umbach, D.M., Terry, P., and Taylor, J.A. (2004). Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics.*, 20(10):1,638-1,640. <https://doi.org/10.1093/bioinformatics/bth098>
- Lin, J.H., and Haug, P.J. (2008). Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of biomedical informatics.*, 41(1):1-14. <https://doi.org/10.1016/j.jbi.2007.06.001>
- Lin, S.W., and Chen, S.C. (2009). PSOLDA: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis. *Applied Soft Computing.*, 9(3):1,008-1,015. <https://doi.org/10.1016/j.asoc.2009.01.001>
- Liu, B., Cui, Q., Jiang, T., and Ma, S. (2004a). A combinational feature selection and ensemble neural network method for classification of gene expression data. *BMC bioinformatics.*, 5(1):1-12.
- Liu, H., and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering.*, 17(4):491-502. <https://doi.org/10.1109/TKDE.2005.66>
- Liu, H., Motoda, H., and Yu, L. (2004b). A selective sampling approach to active feature selection. *Artificial Intelligence.*, 159(1-2):49-74. <https://doi.org/10.1016/j.artint.2004.05.009>
- Liu, J., Chen, C., Liu, Z., Jermisittiparsert, K., and Ghadimi, N. (2020). An IGDT-based risk-involved optimal bidding strategy for hydrogen storage-based intelligent parking lot of electric vehicles. *Journal of Energy Storage.*, 27:101057. <https://doi.org/10.1016/j.est.2019.101057>
- Liu, L., and Liu, J. (2020). Reconstructing gene regulatory networks via memetic algorithm and LASSO based on recurrent neural networks. *Soft Computing.*, 24(6):4,205-4,221. <https://doi.org/10.1007/s00500-019-04185-y>
- Liu, X., and Fu, H. (2014). PSO-based support vector machine with Cuckoo search technique for clinical disease diagnoses. *The Scientific World Journal*, 2014. <https://doi.org/10.1155/2014/548483>
- Liu, Y., and Zheng, Y.F. (2006). FS_SFS: A novel feature selection method for support vector machines. *Pattern*

- recognition., 39(7):1,333-1,345. <https://doi.org/10.1016/j.patcog.2005.10.006>
- Luo, P., Tian, L.P., Ruan, J., and Wu, F. (2017). Disease gene prediction by integrating PPI networks, clinical RNA-Seq data and OMIM data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
- Mehrpooya, M., Ghadimi, N., Marefati, M., and Ghorbanian, S.A. (2021). Numerical investigation of a new combined energy system includes parabolic dish solar collector, Stirling engine and thermoelectric device. *International Journal of Energy Research*. <https://doi.org/10.1002/er.6891>
- Miller, J.F. (2019). Cartesian genetic programming: its status and future. *Genetic Programming and Evolvable Machines*, 1-40. <https://doi.org/10.1007/s10710-019-09360-6>
- Molina, L.C., Belanche, L., and Nebot, À. (2002). Feature selection algorithms: A survey and experimental evaluation. In *2002 IEEE International Conference on Data Mining*, 2002. P. 306-313. <https://doi.org/10.1109/ICDM.2002.1183917>
- Nahato, K.B., Harichandran, K.N., and Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. *Computational and mathematical methods in medicine*, 2015. <https://doi.org/10.1155/2015/460189>
- Nilashi, M., Bin Ibrahim, O., Mardani, A., Ahani, A., and Jusoh, A. (2016). A soft computing approach for diabetes disease classification. *Health Informatics Journal*, 1460458216675500. <https://doi.org/10.1177/1460458216675500>
- Ozcift, A. (2012). SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. *Journal of medical systems*., 36(4):2,141-2,147. <https://doi.org/10.1007/s10916-011-9678-1>
- Özçift, A., and Gülsen, A. (2013). Genetic algorithm wrapped Bayesian network feature selection applied to differential diagnosis of erythematous-squamous diseases. *Digital Signal Processing*., 23(1):230-237. <https://doi.org/10.1016/j.dsp.2012.07.008>
- Pahikkala, T., Okser, S., Airola, A., Salakoski, T., and Aittokallio, T. (2012). Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms for Molecular Biology*., 7(1):11. <https://doi.org/10.1186/1748-7188-7-11>
- Pavlenko, T. (2003). On feature selection, curse-of-dimensionality and error probability in discriminant analysis. *Journal of Statistical Planning and Inference*., 115(2):565-584. [https://doi.org/10.1016/S0378-3758\(02\)00166-0](https://doi.org/10.1016/S0378-3758(02)00166-0)
- Peng, Y., Wu, Z., and Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*., 43(1):15-23. <https://doi.org/10.1016/j.jbi.2009.07.008>
- Plant, C., Teipel, S.J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., and Ewers, M. (2010). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *Neuroimage*., 50(1):162-174. <https://doi.org/10.1016/j.neuroimage.2009.11.046>
- Polat, K., and Güneş, S. (2007). A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. *Computer methods and programs in biomedicine*., 88(2):164-174. <https://doi.org/10.1016/j.cmpb.2007.07.013>
- Pollard, K.S., Dudoit, S., and van der Laan, M.J. (2005). Multiple testing procedures: the multtest package and applications to genomics. In *Bioinformatics and computational biology solutions using R and bioconductor*. Springer, New York, NY. P. 249-271. https://doi.org/10.1007/0-387-29362-0_15
- Rawat, K., and Burse, K. (2013). A soft computing genetic-neuro fuzzy approach for data mining and its application to medical diagnosis. *Int J Eng Adv Technol*., 3:409-411.
- Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., and Jain, A.K. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on evolutionary computation*., 4(2):164-171. <https://doi.org/10.1109/4235.850656>
- Raza, K. (2017). Formal concept analysis for knowledge discovery from biological data. *International journal of data mining and bioinformatics*., 18(4):281-300. <https://doi.org/10.1504/IJDMB.2017.088138>
- Raza, K. (2019). Improving the Prediction Accuracy of Heart Disease with Ensemble Learning and Majority Voting Rule. In *Advances in Ubiquitous Sensing Applications for Healthcare, U-Healthcare Monitoring Systems: Design and Applications*, Academic Press, Elsevier, 179-196.138 <https://doi.org/10.1016/B978-0-12-815370-3.00008-6>
- Raza, K. and Hasan, A.N. (2015). A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction Based on Microarray Data. *International Journal of Bioinformatics Research and Applications*, Inderscience., 11(5):397-416. <https://doi.org/10.1504/IJBRA.2015.071940>
- Raza, K., and Parveen, R. (2013). Soft computing approach for modeling genetic regulatory networks. *Advances in computing and information technology*, P. 1-11. https://doi.org/10.1007/978-3-642-31600-5_1
- Roffo, G. (2016). Feature selection library (MATLAB toolbox). arXiv preprint arXiv:1607.01327.
- Romalt, A.A., and Kumar, R.M.S. (2020). An analysis on feature selection methods, clustering and classification used in heart disease prediction-A machine learning approach. *Journal of Critical Reviews*., 7(6):138-142. <https://doi.org/10.31838/jcr.07.06.27>
- Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*., 23(19):2,507-2,517. <https://doi.org/10.1093/bioinformatics/btm344>
- Seera, M., and Lim, C.P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications*., 41(5):2,239-2,249. <https://doi.org/10.1016/j.eswa.2013.09.022>
- Senliol, B., Gulgezen, G., Yu, L., and Cataltepe, Z. (2008). Fast Correlation Based Filter (FCBF) with a different search strategy. In *2008 23rd international symposium on computer and information sciences*. IEEE. P.1-4. <https://doi.org/10.1109/ISCIS.2008.4717949>
- Shaban, W.M., Rabie, A.H., Saleh, A.I., and Abo-Elsoud, M.A. (2020). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. *Knowledge-Based Systems*, 205, 106270. <https://doi.org/10.1016/j.knosys.2020.106270>
- Shankar, K., Lakshmanprabu, S.K., Gupta, D., Maseleno, A., and De Albuquerque, V.H.C. (2020). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. *The journal of supercomputing*., 76(2):1128-1143. <https://doi.org/10.1007/s11227-018-2469-4>
- Sivagaminathan, R.K., and Ramakrishnan, S. (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert systems with applications*., 33(1):49-60. <https://doi.org/10.1016/j.eswa.2006.04.010>
- Smyth, G.K., Ritchie, M., Thorne, N., and Wettenhall, J. (2005). LIMMA: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions*

- Using R and Bioconductor. Statistics for Biology and Health. https://doi.org/10.1007/0-387-29362-0_23
- Sudha, M. (2017). Evolutionary and Neural Computing Based Decision Support System for Disease Diagnosis from Clinical Data Sets in Medical Practice. *Journal of medical systems.*, 41(11):178. <https://doi.org/10.1007/s10916-017-0823-3>
- Suk, H.I., Lee, S.W., Shen, D., and Alzheimer's Disease Neuroimaging Initiative. (2016). Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Structure and Function.*, 221(5):2,569-2,587. <https://doi.org/10.1007/s00429-015-1059-y>
- Sun, L., Xu, J.C., Wang, W., and Yin, Y. (2016). Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification. *Genetics and molecular research: GMR*, 15(3). <https://doi.org/10.4238/gmr.15038990>
- Thangavel, K., and Pethalakshmi, A. (2009). Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing.*, 9(1):1-12. <https://doi.org/10.1016/j.asoc.2008.05.006>
- Thévenot, E.A. (2016). ropis: PCA, PLS (-DA) and OPLS (-DA) for multivariate analysis and feature selection of omics data.
- Too, J., and Mirjalili, S. (2021). A hyper learning binary dragonfly algorithm for feature selection: A COVID-19 case study. *Knowledge-Based Systems*, 212, 106553. <https://doi.org/10.1016/j.knosys.2020.106553>
- Trevino, V., and Falciani, F. (2006). GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics.*, 22(9):1,154-1,156. <https://doi.org/10.1093/bioinformatics/btl074>
- Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences.*, 98(9):5,116-5,121. <https://doi.org/10.1073/pnas.091062498>
- Vergara, J.R., and Estévez, P.A. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications.*, 24(1):175-186. <https://doi.org/10.1007/s00521-013-1368-0>
- Wang, F., Xu, J., and Li, L. (2014, October). A novel rough set reduct algorithm to feature selection based on artificial fish swarm algorithm. In *International Conference in Swarm Intelligence*. Springer, Cham. P. 24-33. https://doi.org/10.1007/978-3-319-11897-0_4
- Xing, E.P., and Karp, R.M. (2001). CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics.*, 17(1):S306-S315. https://doi.org/10.1093/bioinformatics/17.suppl_1.S306
- Xue, B., Zhang, M., Browne, W.N., and Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation.*, 20(4):606-626. <https://doi.org/10.1109/TEVC.2015.2504420>
- Ye, H., Jin, G., Fei, W., and Ghadimi, N. (2020). High step-up interleaved dc/dc converter with high efficiency. *Energy sources, Part A: recovery, utilization, and environmental effects*, P. 1-20. <https://doi.org/10.1080/15567036.2020.1716111>
- Yin, L., Ge, Y., Xiao, K., Wang, X., and Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing.*, 105:3-11. <https://doi.org/10.1016/j.neucom.2012.04.039>
- Zeng, N., Qiu, H., Wang, Z., Liu, W., Zhang, H., and Li, Y. (2018). A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease. *Neurocomputing.* <https://doi.org/10.1016/j.neucom.2018.09.001>
- Zhang, L., Chen, J., Gao, C., Liu, C., and Xu, K. (2018). An efficient model for auxiliary diagnosis of hepatocellular carcinoma based on gene expression programming. *Medical and biological engineering and computing*, P.1-9. <https://doi.org/10.1007/s11517-018-1811-6>
- Zhao, M., Fu, C., Ji, L., Tang, K., and Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. *Expert Systems with Applications.*, 38(5):5,197-5,204. <https://doi.org/10.1016/j.eswa.2010.10.041>
- Zomaya, A.Y. (2013). Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*, 23, 333. <https://doi.org/10.1002/9781118617151.ch14>