Tadist *et al. J Big Data* (2019) 6:79
https://doi.org/10.1186/s40537-019-0241-0

Journal of Big Data

SURVEY PAPER Open Access

Feature selection methods and genomic big data: a systematic review

Khawla Tadist[1*], Said Najah[2], Nikola S. Nikolov[3], Fatiha Mrabti[4] and Azeddine Zahi[2]

*Correspondence:
khawla.tadist@usmba.ac.ma
[1] Laboratory of Signals, Systems and Components, Laboratory of Intelligent Systems and Applications, Faculty of Sciences and Technologies, Sidi Mohammed Ben Abdellah University, Fez, Morocco
Full list of author information is available at the end of the article

Abstract

In the era of accelerating growth of genomic data, feature-selection techniques are believed to become a game changer that can help substantially reduce the complexity of the data, thus making it easier to analyze and translate it into useful information. It is expected that within the next decade, researchers will head towards analyzing the genomes of all living creatures making genomics the main generator of data. Feature selection techniques are believed to become a game changer that can help substantially reduce the complexity of genomic data, thus making it easier to analyze it and translating it into useful information. With the absence of a thorough investigation of the field, it is almost impossible for researchers to get an idea of how their work relates to existing studies as well as how it contributes to the research community. In this paper, we present a systematic and structured literature review of the feature-selection techniques used in studies related to big genomic data analytics.

Keywords: Systematic review, Mapping process, Genomic big data, Feature selection

Introduction

With the advance of computational techniques, the amount of genomic data has risen exponentially, with a rapid rate [1] making it hard to utilize such data in the medical field without appropriate pre-processing, which in turn leads to more complexity and veracity issues [2] eventually creating multiple complications such as storage, analysis, privacy and security. Therefore, genomic data may look easy to handle in terms of its volume, but it actually requires quite a complicated process due to the complexity, heterogeneity and hybridity of its features. This process is entitled knowledge discovery process [3]:

- *Data recording* Includes the different challenges and tools regarding the capture and storage of data.
- *Data pre-processing* Which includes all the operations of cleaning and appropriation of the captured data to the ready to analyze form in order to optimize the analysis step.
- *Data analysis* The task of evaluating data using different algorithms following a logical reasoning to examine each component of the data provided, with the aim of dispensing insightful outcomes.

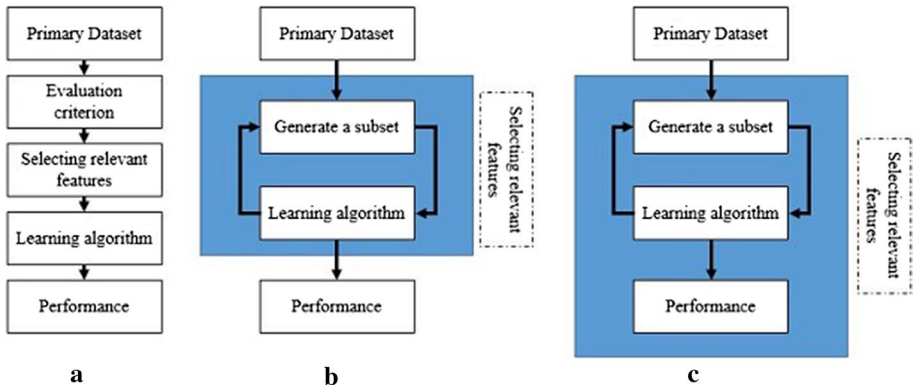

Fig. 1 Basic feature selection methods. The figure shows the three main types of feature selection, filter, wrapper and embedded methods as well as the process in which data passes from initiation to completion of selection in each one

- *Data visualization and interpretation* The step involving the effective knowledge representation using different methods in order to determine the significance and importance of the findings.

The main goal in genomics has primarily been to sequence genomes of all living creature in order to analyze and understand the remaining secrets of the human body and make it possible to detect causes for several genetic diseases. The focus now has evolved from how to sequence the data to how to get use out of the already sequenced data. The multiple challenges that genomic data presents call for the necessity of building a strong model for the preprocessing step. It is compulsory to deal with these challenges in order to allow the decreasing of the volume and complexity by choosing only the most relevant features using feature selection techniques. The preprocessing step is the foundation stone for the analysis accuracy. Even with small databases, genomic data triggers several challenges, such as huge complexity as well as multiplicity of features and attributes, meaning that an appropriate processing step is very critical and needed in order to conduct to perform a high-quality analysis [4]. One of the goals of the preprocessing step is to reduce the dimensionality and the complexity of a dataset, which is accomplished by feature selection. There are mainly six types of feature selection methods. The first three basic methods are (see Fig. 1):

- *Filters* Filter methods are a preprocessing step that is independent of a subsequent learning algorithms. They use independent techniques to select features. The set of features is chosen by an evaluation criterion, or a score to assess the degree of relevance of each characteristics to a target variable [5].
- *Wrappers* Wrappers are feature selection methods that evaluate a subset of characteristics by the accuracy of a predictive model trained along with them. The evaluation is done using a classifier that estimates the relevance of a given subset of characteristics. This type of methods has given evidence to be efficient yet computationally expensive which makes it not very popular [6].
- *Embedded* Combine the qualities of filter and wrapper methods. As the Filter methods have shown to be faster yet not very efficient while the Wrapper methods

are more effective but very computationally expensive especially with big datasets, a solution that combines the advantages of both methods was needed.

Other types of feature selection methods have been identified and praised in the literature. Those types are usually based on the basic three types mentioned above.

- *Hybrid* Methods that apply multiple conjunct primary feature selection methods consecutively [7].
- *Ensemble* Use an aggregate of feature subsets of diverse base classifiers. It consists of the use of different feature subsets [8].
- *Integrative* Integrate external knowledge for feature selection [9].

As a matter of fact, this interest in big genomic data analytics has grown noticeably resulting in a huge amount of publications. The scanning and review of these publications is necessary in order to place a researcher's personal study into the field. Our study presents a systematic mapping of the publications related to the application of feature selection methods in big genomic data analysis using the mapping process suggested method [10] followed by a review of the more relatable publications to our field of study.

The importance of the role of feature selection methods for the processing cycle in big data, and especially genomic big data, is becoming more and more apparent. Many researchers have presented different reviews and surveys of feature selection methods and their role in augmenting results quality.

Vergara et al. [11] highlight and explain the problems that alarm the need for feature selection methods, they offer a state-of-the-art of feature selections methods with an implementation of mutual information feature selection framework. In [12] a set of feature selection methods and classification methods are presented by Li et al. and Mitsunori Ogihara. along with experimental implementations using gene expression datasets. Wang et al. [13] present a survey of feature selection techniques and their applications in big data analysis in the field of bioinformatics offering a new categorization of the feature selection techniques.

In this work, and in contrast to the previously presented related works, that present a classical version of a review paper, we focus on genomic data by following a systematic approach of reviewing the existing feature selection methods specifically in the genomics with the view of helping researchers build a comprehensive perception of the best performing feature selection methods specifically for genomic big data.

The remainder of this paper is organized as follows. In "The mapping process" section, we analyse the different steps of the followed systematic mapping research methodology. "Mapping results and discussion" section highlights the main findings of the mapping process along with the discussion of these findings. In "Review of results" section, we provide a review of the systematic mapping resulting papers. In "Validity considerations" section, we present the validity considerations that were used in the research process. We conclude and share our perspectives and future work in "Conclusion and further work" section.

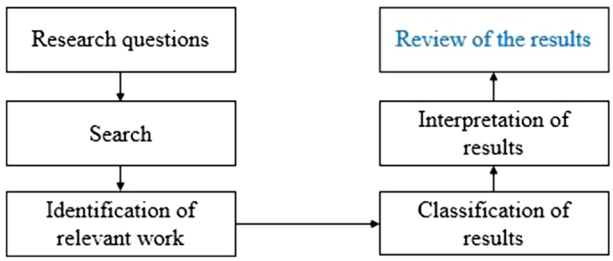

Fig. 2 Adopted systematic mapping process. The figure depicts the systematic mapping and reviewing process adopted in this study, which is proposed by Petersen et al.

The mapping process

In this paper, we employ the systematic mapping process proposed by Petersen et al. [10] (see Fig. 2) with the objective of identifying the most relevant studies that relate to feature selection applied to big genomic data. The main goal of the mapping step is to eventually conduct a review of the mapping step's resulting papers.

Initially, we define the research questions that help shape our study. A query search using key words is run throughout different prominent digital databases, ACM, IEEE Xplore, Science Direct, and Scopus, resulting in an immense number of publications. The next step is the screening of results, which allows to consider only the publications that are centered around our three keywords that are 'Feature Selection', 'Big Data' and 'Genomics'. The last mapping step consists of the classification of the results according to several criteria displayed bellow. At last, a review of the most pertinent works is presented.

Research questions

Research questions are a foundation step contributing in the success of a mapping study. Choosing research questions should be done carefully as it makes or breaks the study. For more accuracy, we have decided to categorize our questions into three types: Guiding Questions (GQ), Categorization Question (CQ), and Discussion Questions (DQ).

The first category is the Guiding Question, which presents a definite and clear expression of the area of concern. For this study, the Guiding Question is expressed as following, "GQ: What type of feature selection techniques are being employed in solving big data problems in genomics?" This question serves as the guiding question of our study. Its purpose is the orientation of the search since the search query is deduced from it. In this study, the articles that englobe the three main parts of this question were taken under consideration, i.e. 'Feature Selection', 'Big Data' and 'Genomics'.

Second questions are Categorization Questions, which may be used in the step of identification of relevant contributions along with the rest of the classification criteria. CQs: What are the categories of feature selection methods according to the tyoe of:

- Research is being conducted in the paper?
- Contribution was proposed in the paper?
- Use in Bioinformatics?
- Data mining: predictive or descriptive?
- Predictive/descriptive modelling?

Table 1 Keywords and synonyms in the search query

Terms	Synonyms list
Feature selection	Variable selection, dimensionality reduction
Big data	Multi-dimensional data, high-dimensional data, Hadoop, MapReduce, Spark
Genomics	Genetics, bioinformatics, micro-array data

CQs are important in a sense that they can be used to decide whether the paper is worth being taken under consideration or not. Along with the inclusion and exclusion criteria, these questions provide enough information for the identification of relevant work.

The last category is the Discussion Question (DQ) that leverages the analytical and critical review of the selected contributions: DQ: What are the feature selection methods and techniques previously employed in big genomic data analytics?

Query search

To conduct this study, we decided to focus on the previously proposed contributions in feature selection techniques that were proposed for big genomic data. Three main terms were selected in order to form an appropriate query for the search. A list of the synonyms of the three terms is also considered in order not to omit any publication that could potentially be relevant (see Table 1).

The list of synonyms help broaden the circle of search, the more terms the query has the higher the chances of not neglecting a relevant work get. In this mapping study, we use dimensionality reduction as a synonym to feature selection although the process of dimensionality reduction actually consists of two sub tasks. The first one is the feature extraction, one important step among the analysis process in any field [14], which involves transforming or projecting a space composing of many dimensions into a space of fewer dimensions and the second task is feature selection which is the process of selecting only relevant and non redundant features. The reason behind using dimentionality reduction as a synonym to feature selection is not to disgard significant papers where the authors might have fused the two tasks or did not clearly state the type of the sub task used. Forming the list of keywords and their synonyms is the helping step for creating the query for the primary search step:

(Feature Selection OR Variable Selection OR Dimensionality Reduction)

AND

(Big Data OR Multi-dimensional data OR High-dimensional data OR Hadoop OR MapReduce OR Spark)

AND

(Genomics OR Genetics OR Bioinformatics OR Micro-array data).

The search query is performed in four of the best assessed digital repositories that prove to respect the worthiness parameters [15]. and was enriched with as many terms related to our study as possible. The primary resulted in a large list of publications, which the length varies depending on each repositories criteria of search (see Table 2).

Table 2 Results of primary search

Repositories	Number of publications
ACM	216,222
IEEE Xplore	38,268
Science Direct	46,180
Scopus	17,600

The large number of publications resulting from performing the query search calls for the need to a thorough selection of papers. This selection is evidently not random, it relies on previously chosen identification criteria depending on the field of the study, as well as other classification standards that meet the expected outcomes of our study.

Identification of relevant work

In order to identify relevant work, several criteria have to be chosen carefully. There are two types of criteria, inclusion and exclusion criteria run through the research question results.

Inclusion criteria

- The reputation of the academic source, such as a journal or conference,
- Articles referenced in one of the articles considered and related to the subject.

Only the papers that were presented in the most prominent journals and conferences were taken under consideration, for higher accuracy, we check the list of the references for the relevant work.

Exclusion criteria

- Delete publications that do not contain the term 'Feature Selection', or any of its synonyms, in the title, summary, or metadata section of the document.
- Delete publications that do not contain the term 'Big Data', or any of its synonyms, in the title, summary, or metadata section of the document.
- Delete documents that only refer to terms without them being a subject of the study.
- Cast aside publications that do not present a strong study that involves the three terms by examining the introduction, conclusion and results sections of each publication.

The finally selected publications need to present a significant work that focuses on all three main terms of the study or else it is not included in our study.

Classification characteristics

The selection relevant studies, after applying the inclusion and exclusion criteria, are classified and categorized according to many characteristics. The first one is the categorization question defined while framing the questions followed by the type of research and the type of contribution and lastly the type of analytics (see Table 3).

Table 3 Publications classification characteristics

Classification characteristics	Types of research	Types of contribution	Types of analytics
Type of characteristics	Evaluation	Architecture	Predictive
	Experience	Framework	Descriptive
	Opinion	Methodology	
	Philosophical	Model	
	Solution	Platform	
	Solution	Process	
	Solution	Theory	
	Solution	Tool	

There are various types of research contributions and the focus in each one differs according to the field of study. We find that in the field of genomics, researchers focus on proposing philosophical, evaluation and solution contributions. They either present a theoretical point of view about a set of existing methods or present a new methodology to solve a recurrent problematic [16].

Types of research

- *Validation research* Research that presents a thorough investigation of a solution that is previously proposed.
- *Experience research* Study where the researcher proposes the steps of an experimental study and presents experimental results.
- *Opinion research* A personal subjective opinion of the researcher focusing on a certain method compared to other related works.
- *Philosophical research* Research that analyses a certain problem on a theoretical level.
- *Solution research* A presented solution to a certain problem supported by experiments and proof of validity.

Types of contribution

- *Architecture* A solution that is constructed of multiple components working together for better results.
- *Framework* A potentially extensible combination of various libraries that solve a certain problem.
- *Methodology* A contribution to the methods for solving a certain computational issue.
- *Model* Presentation of predictive/descriptive models trained for solving particular problems.
- *Platform* A combination of hardware and software solutions enabling applications to run.
- *Process* Data-processing workflows proposed for solving a particular problem.
- *Theory* Philosophical guidance towards solving a certain problem.
- *Tool* Well-defined software utilities addressing a subset of a bigger problem.

Table 4 Results of the identification of relevant work step

Repositories	Number of publications
ACM	1
IEEE Xplore	9
Science Direct	19
Scopus	2

Types of analysis

- *Predictive analysis* An analytical study of current data with the aim of making predictions about future outcomes.
- *Descriptive analysis* An analytical description of the basic features of the dataset in a study that provides simple summaries about a sample.

Mapping results and discussion

The following section presents the outcomes of the step of identification of relevant works in the mapping process highlighted in "The mapping process" section. The difference between the number of publications in each repository is drastic. The reason behind this could be explained by the diversity of the criteria of each search engine (see Table 4).

The ACM search engine is more likely to consider each word on its own during the search and present all the possible articles that contain the word, which explains the enormous difference between the number of the papers resulting from the primary search and the ones that are consequent of the mapping process. The other repositories present a narrower number of publications, which could be explained by the fact that they use more precise and to the point search engines.

After applying the different criteria of inclusion and exclusion, only the most relevant to the field of interest papers are kept. We also choose for reviewing purposes not to include philosophical studies.

Feature selection methods are an important key to the analysis of genomic big data, which calls for the need to more innovative methods and algorithms. It is noticeable that the most researchers in this field offer new innovative solutions, or evaluations of already existing solutions, supported by strong proof and experiments (see Fig. 3). Those two types are followed by validation studies that verify and test with previously proposed solutions. It is also clear that with the advance of years, the number of publications considering more solutions and evaluation paper have gone higher (see Fig. 4).

The interest in the field has grown exponentially both in conferences and journals as we can see in Fig 5. The most noticeable contributions are proposed methodologies that offer new implementations of algorithms. In this field and for the last decade, there are more publications in journals than there are in conferences. Different architectures, frameworks and tools are proposed as well as solutions to the problem of feature selection in big genomic data analytics, yet the methodologies gain the lion's share among the proposed contributions, followed by frameworks and architectures (see Fig. 6).

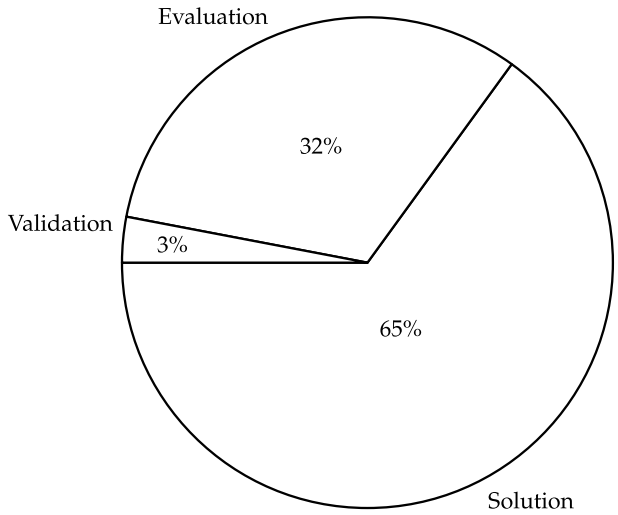

Fig. 3 Existing studies of feature selection methods in big genomic data analytics. The figure shows the results of the comparison between types of already existing studies, which shows that most research is oriented towards presenting solutions or evaluations of already existing solutions, supported by strong proof and experiments

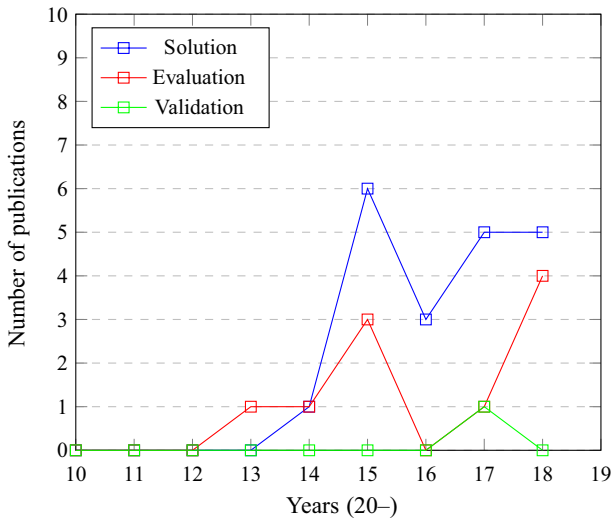

Fig. 4 Distribution of types of research throughout the last ten years. The figure depicts the distribution of the proposed research within the last 10 years, taking into account the type of these studies, solution, evaluation or validation

The goal of data analysis in the medical field is usually predicting diseases with the aim of prevention. When it comes to feature selection methods, the majority of the proposed solutions are part of the preprocessing step in a predictive analytics study, which explains why the publications concerned with predictive analytics outnumber dramatically the ones concerned with descriptive analytics as seen in Fig. 7.

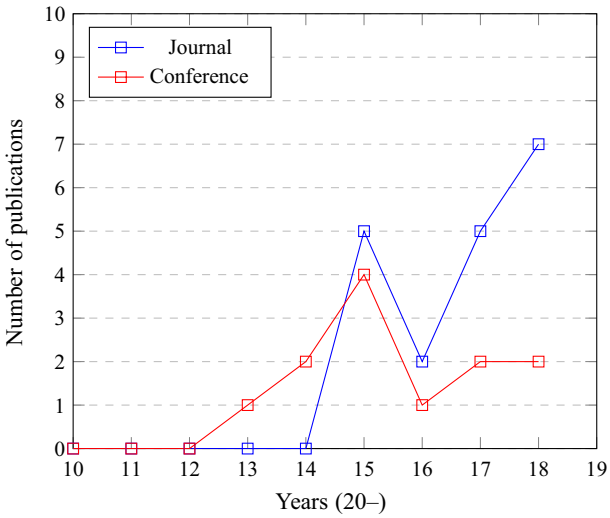

Fig. 5 Year-on-year publication growth in journals and conferences. The figure shows the distribtion of the proposed methods in journals and conferences during the last 10 years

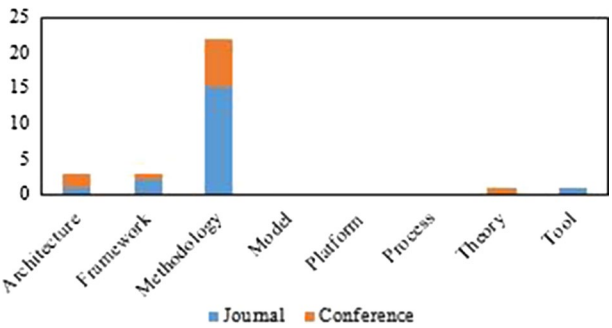

Fig. 6 Distribution of research contributions in journals and conferences. The figure shows the distribution of the research contributions on journals and conferences

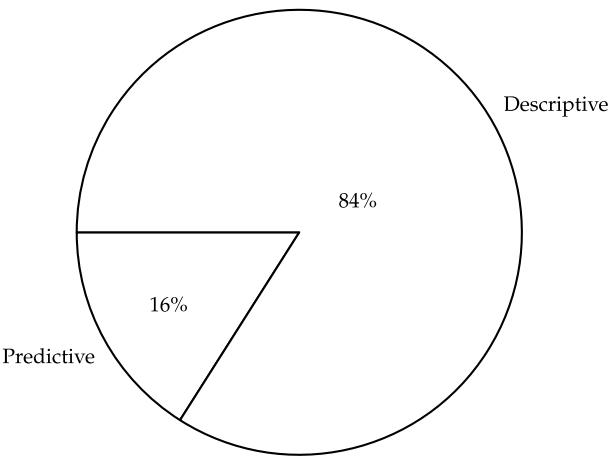

Fig. 7 Types of analysis. The figure depicts the percentage of the number of contributions that lend within the predictive and the descriptive types of analysis

Review of results

DQ: What are the feature selection methods and techniques previously employed in big genomic data analytics?

The crucial role played by the feature selection step has led many researchers to innovate and find different approaches to address this issue. The rationale behind the discussion question is to review and discuss those contributions existing within the resulting papers of the systematic mapping process (Table 5). One distinctive attempt to display an opinion research based on well-known approaches in feature selection applied to digital genetics in order to enhance machine intelligence is found on [17] by Muneshwara et al. In their paper, Muneshwara et al. do not focus on a single type of feature selection method, paradoxically, however, the rest of the contributions display diverse solutions that can be categorized according to the six types of feature selection methods.

Filter methods

The initial feature selection type is the filter methods, in which the algorithm selecting relevant and non-redundant features in the data set is actually independent of the used classifier. Many bioinformatics researchers have shown interest in this particular type of feature selection methods due to the simplicity of its implementation, its low computational cost and its speed. Yang et al. [18] present experimental results of the multivariate (mRMR) feature selection algorithm on five real datasets. The algorithm selects features that have maximal statistical dependency based on mutual information. It considers relevant features and redundant features simultaneously. In another scope, Tsamardinos et al. [19] dispense an algorithm for feature selection in big data settings that can combine local logistic regression coefficients to global models. The algorithm is tested, with Single Nucleotide Polymorphisms (SNP) dataset, against the global logistic regression models produced by Apache MLlib[1] and shows better performance in number of selected features, and predictive performance.

Wrapper methods

Although filter methods are easier to implement, wrapper methods are advantageous for providing better performance by including classification performance of the used classifier, such as accuracy, within the evaluation of the feature selection algorithm. In [20], He et al. present a wrapper feature selection solution for the prediction of a genetic trait, which can be seen as an extension of minimum redundancy maximum relevance (mRMR) feature selection in a transductive manner. Then, using real data they show evidence that their wrapper feature selection leads to higher predictive accuracy than mRMR. On the other hand, analysis of gut microbiota in relation to mental disease (specifically schizophrenia) is the focus of the study in [21], where Shen et al. conduct several experiments using the Boruta feature selection algorithm followed by a random forest classifier are reported. Sun et al., in [22], introduce a new feature selection algorithm for internet of things (IoT) information processing. This method is based on the maximal information coefficient (MIC), allowing to capture different types of correlations

[1] https://spark.apache.org/mllib/.

Table 5 Feature selection methods in big genomic data analytics

Refs	App in genomics	Algorithm	Datasets	Evaluation methods	Technologies	Advantages	Disadvantages	Big data addressed			Type of feature selection					
								Vo	Va	Ve	F	W	Em	H	En	I
[17]	Sorting genomes	–	No datasets	–	–	–	–	–	–	Y	–	–	–	–	–	–
[18]	Classification	mRMR	Colorectal, liverM, pancreatic, central nervous system (CNS), leukemia data	Cross-validation	–	Good classification accuracy	–	Y	Y	–	Y	–	–	–	–	–
[19]	Prediction	PFBP	Nucleotide polymorphism (SNP) data	Bootstrapping	MapReduce	Reduces time complexity with better accuracy parallelized	–	Y	Y	–	Y	–	–	–	–	–
[20]	Genetic trait prediction	MINT	Real data: maize data rice data pine data	Cross-validation	–	Reduces time complexity	–	Y	Y	–	–	Y	–	–	–	–
[21]	Prediction	Boruta Random Forest	Next- generation sequencing laboratory of Novogene Bioinformatics Institute, Beijing, China,	Bootstrapping	NetBeans	Good prediction accuracy	Small sample size	Y	Y	–	–	Y	–	–	–	–
[22]	Prediction	MIMIC FS	ASU datasets	Cross-validation	Weka	Good performance	–	Y	Y	–	–	Y	–	–	-	–
[23]	Marker selection	FIFS	Single nucleotide polymorphism (SNP)	Train and test	–	Huge rate of success	Not parallelized	Y	Y	–	–	Y	–	–	–	–
[24]	Binning for prediction	Random forest Naïve Bayes	Generated datasets	Train and test	–	Dataset presents better prediction	–	Y	Y	–	–	–	Y	–	–	–
[25]	Classification predicting disease	SVEGA	Breast cancer dataset Kent ridge biomedical repository	TPR/FPR	–	Classification accuracy rate	Not parallelized	Y	Y	-	-	-	Y	-	-	-
[26]	Classification prediction	SVM	Kent Ridge Bio-medical dataSet Repository and National center of Biotechnology Information	ANOVA	Hadoop MapReduce	Good accuracy rate	–	Y	Y	–	–	–	Y	–	–	–
[27]	Classification prediction	K-nearest neighbor	National Center of Biotechnology Information NCBI GEO	Cross-validation	Hadoop MapReduce	Reduces time complexity Parallelized	–	Y	Y	–	–	–	–	–	–	Y

Table 5 (continued)

Refs	App in genomics	Algorithm	Datasets	Evaluation methods	Technologies	Advantages	Disadvantages	Big data addressed			Type of feature selection					
								Vo	Va	Ve	F	W	Em	H	En	I
[28]	Identification of gene expression signatures	SVM	20,475 features in 1920 samples, a highdimensional dataset (source not mentioned)	Cross-validation	Weka	Better understanding of the classification	–	Y	Y	–	–	–	Y	–	–	–
[29]	Prediction	Cox-regression	The Cancer Genome Atlas datasets, glioblastoma) and lung adenocarcinoma	Cross-validation	–	Higher true variables rate Better predicting performance Easy-to-implement property	–	Y	Y	–	–	–	Y	–	–	–
[30]	Prediction	mRMR IFS	Genome-wide association studies	Cross-validation	Weka	Good classification performance	Not parallelized	Y	Y	–	–	–	Y	–	–	–
[31]	Prediction	mRMR IFS	UniProtdatabase http://www.uniprot.org	Cross-validation	Weka	High prediction accuracy	Not parallelized	Y	Y	–	–	–	Y	–	–	–
[32]	Classification	ROSEFW-RF	Generated with the ROS technique	Train and test	MapReduce	Parallelized Suitable for large scale data	–	Y	Y	–	–	–	Y	–	–	–
[33]	Genetic association	Screening	GEO database with ID GSE13355 and GSE14905	Cross-validation	–	Good classification accuracy	–	Y	Y	–	–	–	–	Y	–	–
[34]	Classification	Decision Tree Support Vector Machine	UCI machine learning repository http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html	Cross-validation	Weka	Simplicity of implementation Reduces time complexity High accuracy	Expensive computational cost	Y	Y	–	–	–	–	Y	–	–
[35]	Prediction	Pearson Correlation Coefficient (PCC) Information Gain (IG) and ReliefF	Prokaryotic model organism name as *E. coli*, as a real biologic network	Fitness function	–	High speed and prediction accuracy Easily parallelizable	–	Y	Y	–	–	–	–	Y	–	–

Table 5 (continued)

Refs	App in genomics	Algorithm	Datasets	Evaluation methods	Technologies	Advantages	Disadvantages	Big data addressed			Type of feature selection					
								Vo	Va	Ve	F	W	Em	H	En	I
[36]	Classification	SVM	Sentiment classification of on-line reviews using data collected from amazon, imdb, and yelp. Cancer classification based on gene ex-pressions for leukemia, prostate cancer, and lung cancer	Hold-out validation	–	Simplicity and low error rates	Lack of scalability Not parallelized	Y	Y	–	–	–	–	Y	–	–
[37]	Classification	SVM	Breast cancer, colorectal adenocarcinoma, head and neck squamous cell carcinoma, kidney renal clear cell carcinoma, ovarian cancer http://www.cbio.mskcc.org/cancergenomics/pancan	Cross-validation	Weka	Optimal classification	–	Y	Y	–	–	–	Y	–	–	–
[38]	Clustering	Different clustering algorithms	Exome dataset of Brugada syndrome (BrS)	–	–	Suitable for high-dimensional genomic big data	No parallel implementation	Y	Y	–	–	–	–	–	Y	–
[39]	Classification next generation sequencing	SVM Random Forest	NCBI Reference sequence database, http://http://www.ncbi.nlm.nih.gov/refseq/	Cross-validation	Hadoop MapReduce	Scalable High classification accuracy	–	Y	Y	–	–	–	–	–	Y	–
[40]	Identification of genetic markers prediction	Sparse Regression	SNP: a database of single nucleotide polymorphisms http://www.alzgene.org	Cross-validation	-	Good accuracy for selection of features	Not always trivial	Y	Y	–	–	–	–	–	–	Y
[41]	Prediction detecting SNP interactions	LogicFS-GPU	Stimulated and real schizophrenia data set	Cross-validation	MapReduce	Parallel design of the algorithm	Expensive computational cost	Y	Y	–	–	–	–	–	–	Y

Table 5 (continued)

Refs	App in genomics	Algorithm	Datasets	Evaluation methods	Technologies	Advantages	Disadvantages	Big data addressed			Type of feature selection					
								Vo	Va	Ve	F	W	Em	H	En	I
[42]	Sequencing	PrefDiv and MGM PC-Stable	Pathway information database Cancer Genome Atlas (TCGA)	Cross-validation	-	Combining two algorithms to inhance accuracy	–	Y	Y	–	–	–	–	–	–	Y
[43]	Prediction	Fireflies and ant colony	PDB Bank dataset Varibench Protein data Lung Cancer data bank Marketing	TPR/FPR	Matlab	High efficiency for feature selection	–	Y	Y	–	–	–	–	–	–	Y
[44]	Classification for prediction	ANOVA and K-Nearest Neighbor	NCBI GEO Leukemia Ovarian Cancer Breast Cancer	ANOVA	MapReduce	Distributed and scalable	–	Y	Y	–	–	–	–	–	–	Y
[45]	Classification for prediction	Decision tree k-nearest-neighbor	Brugada syndrome at Centre for Medical Genetics http://www.uzbrussel.be	Cross-validation	Weka	Good prediction accuracy Good with heterogeneous data	–	Y	Y	–	–	–	–	–	–	Y
[46]	Classification	–	Real-life biomedical data, SNP repository data; mixture models simulation studies	Cross-validation	MapReduce	High classification performance Parallelized	–	Y	Y	–	–	–	–	–	–	Y
[47]	Classification	–	Graph datasets of protein 3D-structures	Cross-validation	MapReduce	Improves prediction accuracy	–	Y	Y	–	–	–	–	–	–	Y

Vo volume, *Va* variety, *Ve* velocity, *F* filter, *W* wrapper, *Em* embedded, *H* hybrid, *En* ensemble, *I* integreted

(–) Refers to a lack of definition in the referenced papers

(Y) Refers to the suitability with the titles

between variables. A new data mining approach, called frequent item feature selection is proposed by Kavakiotis et al. [23], the novelty approach is based on the use of frequent items for the selection of most informative markers from genomic data, relying on two major components, the first being the identification of the most frequent and unique genotypes for each sampled population and the second being the selection of the most informative SNP subsets among these populations.

Embedded methods

Embedded methods work by adding a penalty against complexity to reduce the degree of overfitting or variance of a model by adding more bias. Those methods are different from other feature selection methods in the way that feature selection and learning interact; they do not separate the learning from the feature selection part. In [24], Saghir et al. present an evaluation of a random forest classifier using generated datasets for whole genome shotgun (WGS) sequencing in order to solve binning and classification problems. Diversely, Sasikala et al. propose in [25] a genetic algorithm for feature selection method, called SVEGA to rank genes according to their capability to differentiate the classes. Tests with four classification algorithms demonstrate its ability to reduce features and improve accuracy rate. Alternatively, Kumar et al. in [26] propose a method that includes a diversity of statistical tests for feature selection. Similarly to [27], they use a distributed implementation based on MapReduce on Hadoop in order to reduce execution time. In the same scope, in [28] Zhang et al. apply a novel computational strategy to identify gene expression signatures in three types of hematopoietic cells, where each cell type is represented by its gene expression profile. To achieve this goal, the expression features are analyzed by a combination of a Monte Carlo feature selection (MCFS) algorithm and an optimized SVM classifier method, resulting in a feature list of the relevant gene expression.

Liu et al. developed in [29] two methods SKI-Cox and wLASSO-Cox, respectively, to facilitate variable selection for Cox-regression model using multi-omics data. They propose a new framework that can be useful in building a clinically applicable predictive models, as well as identifying driver genes helping to explaining cancer development, prognosis, and relation to patient-specific outcomes. Within the same scope of embedded methods, Li and Huang [30] attempt to identify characteristic tissue-gene expression patterns through the combination of morningness-associated genetic polymorphisms in a genome-wide association studies (GWAS) data. For this, the authors employ an incremental feature selection method with a dagging classifier, to analyze tissue-gene expression patterns and extract the important ones. Zhou et al. in [31] propose a computational method to predict *N*-formyl methionines (fMet) based on various types of features, including position-specific scoring matrix (PSSM) based conservation scores, amino acid factors, secondary structures, solvent accessibilities and disorder scores. The optimal set of features is extracted using mRMR and incremental feature selection (IFS) methods. On the other hand, in [32] Triguero et al. present random oversampling and evolutionary feature weighting for a random forest (ROSEFW-RF) algorithm, which reportedly deals well with imbalanced class distribution in a large dataset. Prior to building the model, they apply a combination of multiple preprocessing stages, such

as random oversampling, a evolutionary feature weighting. All steps of this approach are run within MapReduce computational framework.

Hybrid methods

Hybrid methods gained an immense popularity due to the fact that they incorporate multiple types of feature selection methods, Filters, Wrappers and Embedded, within the same process. In [33], Wang et al. apply two screening method on a publicly available sample from the Gene Expression Omnibus (GEO) database.[2] The methods help omit redundant genomic pairs that do not help the prediction process and reduce correlation among classifiers in order to improve prediction accuracy. With the aim of speeding up the training time of a support vector machine (SVM) algorithm, Arumugam and Jose in [34] propose an algorithm that utilizes a twofold SVM and applies decision tree as a data filter, to reduce dimensionality. Alternatively, in [35] a framework that incorporate the Pearson correlation coefficient within two different feature selection approaches based on information gain and relief is proposed by Jafari et al. The framework is tested on real biological data showing higher accuracy and speed compared to other state-of-the-art methods. From another perspective, Ghaddar et al. in [36] address the problem of selecting the minimal number of features for a binary classifier. They introduce a new approach for SVM classification and feature selection based on iteratively adjusting a bound on the l1-norm of the classifier vector. Reportedly, the advantage of this approach is its intuitive implementation and computational tractability for applications that contain high dimensional features where the direct application of standard feature selection models is computationally intractable. On the same premises, in [37], Wang et al. employ (MCFS) followed by incremental feature selection (IFS) to identify relevant features that can be used to train an SVM classifier for distinguishing the five types of cancers. The use of MCFS in feature analysis leads to the extraction of 16 decision rules that augment the classification accuracy.

Ensemble methods

Ensemble feature selection methods combine independent feature subsets and could eventually provide better approximation to the optimal subset of features, which made them attract researchers' attention during the last few years. In [38], Farid et al. propose a feature selecting method for ensemble clustering of complex genomic data by combining two traditional clustering algorithms, k-means and similarity-based clustering. They test their model on an exome data set (for Brugada syndrome studies) and compare it with four different clustering methods, showing that their method results to decreasing compactness. Within the same scope of Ensemble methods, Hogan et al. address the problem of Next Generation Sequencing (NGS) data at very large scale in [39] by investigating the effectiveness of parallel ensemble classifiers, principally random forests, to take advantage of the available computational resources. They consider a mix of real and synthetic data.

[2] https://www.ncbi.nlm.nih.gov/geo/.

Itegrative methods

Integrative feature selection methods are considered as an immerging genre, they integrate external data during the process of feature selection. Zhu et al. present in [40] an implementation of a sparse regression algorithm. They integrate an additional regression technique in order to increase the feature selection accuracy. Their algorithm is tested upon a complex (SNP) database and indeed shows better results than other feature selection methods they experimented with. Alternatively, in [41], Altinigneli et al. present a parallelized form of the LogicFS algorithm applied on stimulated datasets and real schizophrenia datasets for predicting SNP interactions and shows a great running time improvement compared to non-parallelized LogicFS. On the other hand, in [42], Raghu et al. present an integrative feature selection method for finding a maximally relevant and diverse gene sets with preferential diversity using an importance score that combines both prior knowledge and data inherent information. On the strength of itegrative feature selection methods, AlFarraj et al. [43] examine the Ant Colony Optimization based feature selection process. They use various datasets such as the Protein Data Bank,[3] VariBench protein data,[4] lung cancer data and, bank marketing data in order to investigate the accuracy of the method, which shows better performance compared to other methods.

Based on MapReduce paradigm, Kumar et al. [44] propose the usage of an ANOVA statistical test for feature selection, followed by training k-nearest neighbors (kNN) classifier for classifying big microarray data. They utilize MapReduce over scalable clusters which also allows the processing time to be reduced. Presented in [45] another research direction that dealing with Rule-based classifier where Farid et al. propose an adaptive rule-based classifier for multi-class classification of biological data, in a human interpretable way. Their classifier combines the random subspace and boosting approaches with an ensemble of decision trees to generate a set of classification rules with the goal of minimizing over-fitting and the impact of, noisy instances and class-imbalance in data. In order to select relevant features, Kumar et al. [27] propose a method that uses various statistical tests, followed by kNN to classify data into cancerous/non-cancerous. The distributed implementation (MapReduce on Hadoop) of these methods reportedly reduces the execution time drastically. Within the same scope, Elsebakhi et al. propose, in [46], a functional networks method for enhancing a large scale machine learning classifier based on propensity score and Newton-Raphson-maximum likelihood optimization. The application of this method on big biomedical data shows that this method outperforms most of the existing state-of-the-art statistical and machine learning methods with regards to performance and execution time. Based on MapReduce, Dhifli et al. [47] propose the scalable and distributed method MR-SimLab to compute pairwise similarities between labels of graph nodes. A comparative study on multiple datasets shows that this method improves predictive accuracy.

[3] https://www.rcsb.org/.

[4] http://structure.bmc.lu.se/VariBench/.

Validity considerations

In order to present a significant review, a very critical process has to be followed. It is preferable to apply a systematic mapping process upon the set of publications in the field of interest. Although it is always preferable to follow a systematic process, before engaging in a functional study, it still cannot present a perfect accuracy and reliability. During this study, we have tried to limit the risks of error yet that does not negate the fact that there are still some threats to the validity of the process.

Digital databases

Since we used a limited number of well-recognized repositories, i.e. ACM, IEEE Xplore, Science Direct and Scopus, to select the initial set of papers, we may have omitted some possibly strong contributions. The decision to neglect other repositories can be justified by the fact that if a paper presents a strong contribution, chances are it would be referenced in one of the initially selected papers, and thus, it would be included in our study after applying the second inclusion criteria.

Research questions

Our research questions were discussed and agreed by the members of the research team. There is a chance that an aspect of interest to other researchers may have been neglected. Although the team welcomed external and internal propositions about what could be an aspect of interest, there is a slight chance that some angles could not have been covered by this study.

Inclusion and exclusion criteria

As inclusion and exclusion criteria can have major impact on the mapping process, they are also agreed on previously by the whole research team. To the best of knowledge, we include all possible synonyms of the key search terms.

Classification accuracy

The labeling of each research, publication and type of analytics proposed was quite difficult, each paper was checked twice in order to verify its categorization. We have tried our hardest to match the conventional agreed upon classification elements in order to upgrade the accuracy of the categorization thus limit the chances of error.

Conclusion and further work

In this paper, focusing on the data preprocessing step, we identify and review the most relevant studies on feature selection methods employed in the analysis of genomic big data. We believe that our work will benefit future studies in genomic data analytics. The review of research literature highlights the strong correlation between the choice of appropriate feature selection methods and the nature of the dataset as well as the type of the study and desired outputs. A wide range of the reviewed papers propose new solutions through offering new methodologies, frameworks, architectures and tools depicting the importance of the usage of feature selection methods while processing genomic big data. In another scope, a considerable amount of papers

offer evaluation and validation tests of previously proposed methodologies and tools. Despite the increasing interest in genomic data analytics, the attention on the preprocessing step remains consistently present. As future work, we aim at contributing to the feature selection methods by proposing a hybrid feature selection method for genomic data and evaluate it within a genomic analytics process.

Abbreviations
mRMR: minimum redundancy maximum relevance; SNP: single nucleotide polymorphisms; IoT: internet of things; MIC: maximal information coefficient; WGS: whole genome shotgun; MCFS: Monte Carlo feature selection; GWAS: genome-wide association studies; fMet: formyl methionines; PSSM: position-specific scoring matrix; GEO: gene expression omnibus; SVM: support vector machine; MCFS: Monte Carlo feature selection; IFS: incremental feature selection; NGS: next generation sequencing; kNN: k-nearest neighbors; CNS: central nervous system; PCC: Pearson correlation coefficient; IG: information gain; BrS: Brugada syndrome.

Acknowledgements
The authors thank the anonymous reviewers for their helpful suggestions and comments

Authors' contributions
All mentioned authors contribute in the elaboration of the paper. All authors read and approved the fnal manuscript.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Laboratory of Signals, Systems and Components, Laboratory of Intelligent Systems and Applications, Faculty of Sciences and Technologies, Sidi Mohammed Ben Abdellah University, Fez, Morocco. [2] Laboratory of Intelligent Systems and Applications, Faculty of Sciences and Technologies, Sidi Mohammed Ben Abdellah University, Fez, Morocco. [3] Department of Computer Science and Information Systems (CSIS) at the University of Limerick, Limerick, Ireland. [4] Laboratory of Signals, Systems and Components Faculty of Sciences and Technologies, Sidi Mohammed Ben Abdellah University, Fez, Morocco.

Appendix

See Table 6.

Table 6 List of the reviewed papers

Publication repository	Type	Title	Year	Classification
IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)	Conference	Minimal-redundancy-maximal-relevance feature selection using different relevance measures for omics data classification [18]	2013	Evaluation
IEEE International Conference on Big Data	Conference	Identification of SNP Interactions Using Data-Parallel Primitives on GPUs [41]	2014	Solution
Elsevier—Procedia Computer Science	Conference	Large Scale Read Classification for Next Generation Sequencing [39]	2014	Evaluation
IEEE International Symposium on Technologies for Homeland Security (HST)	Conference	Big Data Biology-Based Predictive Models Via DNAMetagenomics Binning For WMD Events Applications [24]	2015	Evaluation
IEEE International Congress on Big Data	Conference	Two screening methods for genetic association study with application to psoriasis microarray data sets [33]	2015	Solution
Elsevier—Procedia Computer Science	Conference	A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer [25]	2015	Solution
Elsevier—Procedia Computer Science	Conference	Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor [44]	2015	Evaluation
Elsevier—Knowledge-Based Systems	Journal	Classification of microarray using MapReduce based proximal support vector machine classifier [26]	2015	Solution
Elsevier—Journal of Computational Science	Journal	Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms [46]	2015	Solution
Elsevier—Knowledge-Based Systems	Journal	ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem [32]	2015	Solution
IEEE/ACM Transactions On Computational Biology And Bioinformatics	Journal	MINT: Mutual Information based Transductive Feature election for Genetic Trait Prediction [20]	2015	Solution
IEEE Future Technologies Conference	Conference	A Feature Grouping Method for Ensemble Clustering of High-Dimensional Genomic Big Data [38]	2016	Solution
Elsevier—Expert Systems With Applications	Journal	An adaptive rule-based classifier for mining big biological data [45]	2016	Solution
Elsevier—Journal of Biomedical Informatics	Journal	Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier [27]	2016	Solution
Elsevier—Neurocomputing	Journal	Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method [31]	2016	Evaluation
IEEE Transactions On Big Data Journal	Journal	Low-Rank Graph-Regularized Structured Sparse Regression for Identifying Genetic Biomarkers [40]	2017	Solution
IEEE Digital Genomics to Build a Smart Franchise in Real Time Applications	Conference	Digital Genomics to Build a Smart Franchise in Real Time Applications [17]	2017	Opinion

Table 6 (continued)

Publication repository	Type	Title	Year	Classification
IEEE 33rd International Conference on Data Engineering	Conference	Integrated Theory- and Data-driven Feature Selection in Gene Expression Data Analysis [42]	2017	Solution
Elsevier—Artificial Intelligence In Medicine	Journal	A hybrid framework for reverse engineering of robust Gene Regulatory Networks [35]	2017	Solution
Elsevier—Computers in Biology and Medicine	Journal	FIFS: A data mining method for informative marker selection in high dimensional population genomic data [23]	2017	Solution
Elsevier—Information Systems	Journal	MR-SimLab: Scalable subgraph selection with label similarity for big data [47]	2017	Solution
Elsevier—Methods	Journal	Multi-omics facilitated variable selection in Cox-regression model for cancer prognosis prediction [29]	2017	Evaluation
Springer	Journal	A greedy feature selection algorithm for Big Data of high dimensionality [19]	2018	Solution
Springer—The Natural Computing Applications Forum	Conference	Optimized feature selection algorithm based on fireflies with gravitational ant colony algorithm for big data predictive analytics [43]	2018	Solution
Elsevier—Materials today Proceedings	Conference	Efficient Decision Tree Based Data Selection and Support Vector Machine Classification [34]	2018	Solution
Elsevier—Schizophrenia Research	Journal	Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: A cross-sectional study [21]	2018	Evaluation
Elsevier—BBA-Molecular Basis of Disease	Journal	Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine [28]	2018	Evaluation
Elsevier—BBA-Molecular Basis of Disease	Journal	Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies [30]	2018	Evaluation
Elsevier—Future Generation Computer Systems	Journal	Feature selection for IoT based on maximal information coefficient [22]	2018	Solution
Elsevier—European Journal of Operational Research	Journal	High dimensional data classification and feature selection using support vector machines [36]	2018	Solution
Elsevier—BBA-Molecular Basis of Disease	Journal	Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis [37]	2018	Evaluation

Received: 5 May 2019 Accepted: 8 August 2019
Published online: 27 August 2019

References

1. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. IEEE J Biomed Health Inform. 2015;19(4):1193.
2. West M, Ginsburg GS, Huang AT, Nevins JR. Embracing the complexity of genomic data for personalized medicine. Genome Res. 2006;16(5):559.

3. Chen CP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf Sci. 2014;275:314.
4. Berrar D, Bradbury I, Dubitzky W. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics. 2006;22(10):1245.
5. Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data. 2015;2(1):24.
6. Kushmerick N, Weld DS, Doorenbos R. Wrapper induction for information extraction. Washington: University of Washington; 1997.
7. Naseriparsa M, Bidgoli AM, Varaee T. A hybrid feature selection method to improve performance of a group of classification algorithms. 2014. arXiv preprint arXiv:1403.2372.
8. Tsymbal A, Pechenizkiy M, Cunningham P. Diversity in search strategies for ensemble feature selection. Inf Fusion. 2005;6(1):83.
9. Grasnick B, Perscheid C, Uflacker M. A framework for the automatic combination and evaluation of gene selection methods. In: International conference on practical applications of computational biology & bioinformatics. Berlin: Springer; 2018. p. 166–74.
10. Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic mapping studies in software engineering. Ease. 2008;8:68–77.
11. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Appl. 2014;24(1):175.
12. Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics. 2004;20(15):2429.
13. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: a survey from the search perspective. Methods. 2016;111:21.
14. Kumar S, Zymbler M. A machine learning approach to analyze customer satisfaction from airline tweets. J Big Data. 2019;6(1):62.
15. Houghton B. Trustworthiness: self-assessment of an institutional repository against ISO 16363–2012. D-Lib Mag. 2015;21(3/4):1.
16. O'Donovan P, Leahy K, Bruton K, O'Sullivan DT. Big data in manufacturing: a systematic mapping study. J Big Data. 2015;2(1):20.
17. Muneshwara M, Swetha M, Thungamani M, Anil G. Digital genomics to build a smart franchise in real time applications, In: 2017 international conference on circuit, power and computing technologies (ICCPCT). New York: IEEE; 2017. p. 1–4.
18. Yang J, Zhu Z, He S, Ji Z. Minimal-redundancy-maximal-relevance feature selection using different relevance measures for omics data classification. In: 2013 IEEE symposium on computational intelligence in bioinformatics and computational biology (CIBCB). New York: IEEE; 2013. p. 246–51.
19. Tsamardinos I, Borboudakis G, Katsogridakis P, Pratikakis P, Christophides V. A greedy feature selection algorithm for Big Data of high dimensionality. Mach Learn. 2019;108(2):149–202.
20. He D, Rish I, Haws D, Parida L. Mint: mutual information based transductive feature selection for genetic trait prediction. IEEE/ACM Trans Comput Biol Bioinform. 2016;13(3):578.
21. Shen Y, Xu J, Li Z, Huang Y, Yuan Y, Wang J, Zhang M, Hu S, Liang Y. Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: a cross-sectional study. Schizophr Res. 2018;197:470.
22. Sun G, Li J, Dai J, Song Z, Lang F. Feature selection for IoT based on maximal information coefficient. Future Gener Comput Syst. 2018;89:606.
23. Kavakiotis I, Samaras P, Triantafyllidis A, Vlahavas I. FIFS: a data mining method for informative marker selection in high dimensional population genomic data. Comput Biol Med. 2017;90:146.
24. Saghir H, Megherbi DB. Big data biology-based predictive models via DNA-metagenomics binning for WMD events applications. In: 2015 IEEE international symposium on technologies for homeland security (HST). New York: IEEE; 2015. p. 1–6.
25. Sasikala S, alias Balamurugan SA, Geetha S. A novel feature selection technique for improved survivability diagnosis of breast cancer. Procedia Comput Sci. 2015;50:16.
26. Kumar M, Rath SK. Classification of microarray using MapReduce based proximal support vector machine classifier. Knowl Based Syst. 2015;89:584.
27. Kumar M, Rath NK, Rath SK. Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier. J Biomed Inform. 2016;60:395.
28. Zhang YH, Hu Y, Zhang Y, Hu LD, Kong X. Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine. Biochim Biophys Acta Mol Basis Dis. 2018;1864(6):2255.
29. Liu C, Wang X, Genchev GZ, Lu H. Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine. Methods. 2017;124:100.
30. Li J, Huang T. Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies. Biochim Biophys Acta Mol Basis Dis. 2018;1864(6):2241.
31. Zhou Y, Huang T, Huang G, Zhang N, Kong X, Cai YD. Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method. Neurocomputing. 2016;217:53.
32. Triguero I, del Río S, López V, Bacardit J, Benítez JM, Herrera F. ROSEFW-RF: the winner algorithm for the ECBDL'14 big data competition: an extremely imbalanced big data bioinformatics problem. Knowl Based Syst. 2015;87:69.
33. Wang MH, Tsoi K, Lai X, Chong M, Zee B, Zheng T, Lo SH, Hu I. Two screening methods for genetic association study with application to psoriasis microarray data sets. In: 2015 IEEE international congress on big data. New York: IEEE; 2015. p. 324–6.
34. Arumugam P, Jose P. Efficient decision tree based data selection and support vector machine classification. Mater Today Proc. 2018;5(1):1679.
35. Jafari M, Ghavami B, Sattari V. A hybrid framework for reverse engineering of robust gene regulatory networks. Artif Intell Med. 2017;79:15.

36. Ghaddar B, Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. Eur J Oper Res. 2018;265(3):993.
37. Wang S, Cai Y. Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis. Biochim Biophys Acta Mol Basis Dis. 2018;1864(6):2218.
38. Farid DM, Nowe A, Manderick B. A feature grouping method for ensemble clustering of high-dimensional genomic big data. In: 2016 future technologies conference (FTC). New York: IEEE; 2016. p. 260–8.
39. Hogan JM, Peut T. Large scale read classification for next generation sequencing. Procedia Comput Sci. 2014;29:2003.
40. Zhu X, Suk HI, Huang H, Shen D. Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers. IEEE Trans Big Data. 2017;3(4):405.
41. Altinigneli C, Konten B, Rujescir D, Böhm C, Plant C. Identification of SNP interactions using data-parallel primitives on GPUs. In: 2014 IEEE international conference on big data (Big Data). New York: IEEE; 2014. p. 539–48.
42. Raghu VK, Ge X, Chrysanthis PK, Benos PV Integrated theory-and data-driven feature selection in gene expression data analysis. In: 2017 IEEE 33rd international conference on data engineering (ICDE). New York: IEEE; 2017. p. 1525–32.
43. AlFarraj O, AlZubi A, Tolba A. Optimized feature selection algorithm based on fireflies with gravitational ant colony algorithm for big data predictive analytics. Neural Comput Appl. 2018:1–13.
44. Kumar M, Rath NK, Swain A, Rath SK. Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor. Procedia Comput Sci. 2015;54:301.
45. Farid DM, Al-Mamun MA, Manderick B, Nowe A. An adaptive rule-based classifier for mining big biological data. Expert Syst Appl. 2016;64:305.
46. Elsebakhi E, Lee F, Schendel E, Haque A, Kathireason N, Pathare T, Syed N, Al-Ali R. Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. J Comput Sci. 2015;11:69.
47. Dhifli W, Aridhi S, Nguifo EM. MR-SimLab: scalable subgraph selection with label similarity for big data. Inf Syst. 2017;69:155.

Publisher's Note