Briefings in Bioinformatics, 22(5), 2021, 1–19

https://doi.org/10.1093/bib/bbab128
Method Review

Machine learning approach to gene essentiality prediction: a review

Olufemi Aromolaran, Damilare Aromolaran, Itunuoluwa Isewon and Jelili Oyelade

Corresponding author: Jelili Oyelade, Tel: +2348035755778, E-mail: ola.oyelade@covenantuniversity.edu.ng

Abstract

Essential genes are critical for the growth and survival of any organism. The machine learning approach complements the experimental methods to minimize the resources required for essentiality assays. Previous studies revealed the need to discover relevant features that significantly classify essential genes, improve on the generalizability of prediction models across organisms, and construct a robust gold standard as the class label for the train data to enhance prediction. Findings also show that a significant limitation of the machine learning approach is predicting conditionally essential genes. The essentiality status of a gene can change due to a specific condition of the organism. This review examines various methods applied to essential gene prediction task, their strengths, limitations and the factors responsible for effective computational prediction of essential genes. We discussed categories of features and how they contribute to the classification performance of essentiality prediction models. Five categories of features, namely, gene sequence, protein sequence, network topology, homology and gene ontology-based features, were generated for *Caenorhabditis elegans* to perform a comparative analysis of their essentiality prediction capacity. Gene ontology-based feature category outperformed other categories of features majorly due to its high correlation with the genes' biological functions. However, the topology feature category provided the highest discriminatory power making it more suitable for essentiality prediction. The major limiting factor of machine learning to predict essential genes conditionality is the unavailability of labeled data for interest conditions that can train a classifier. Therefore, cooperative machine learning could further exploit models that can perform well in conditional essentiality predictions.

Short abstract

Identification of essential genes is imperative because it provides an understanding of the core structure and function, accelerating drug targets' discovery, among other functions. Recent studies have applied machine learning to complement the experimental identification of essential genes. However, several factors are limiting the performance of machine learning approaches. This review aims to present the standard procedure and resources available for predicting essential genes in organisms, and also highlight the factors responsible for the current limitation in using machine learning for conditional gene essentiality prediction. The choice of features and ML technique was identified as an important factor to predict essential genes effectively.

Key words: essential genes; essential proteins; feature selection; supervised learning; conditional essentiality; conditionally essential genes

Olufemi Aromoran is a Ph.D. research fellow in the Department of Computer and Information Sciences, Covenant University. His research focuses on applied machine learning and genomics. He is a member of the Nigerian Bioinformatics and Genomics Network.

Damilare Aromolaran is a Ph.D. fellow in the computer science department at Covenant University. Her research focus is on data analytics and data warehousing.

Dr Isewon, Itunuoluwa Marian is a lecturer in the Department of Computer and Information Sciences, Covenant University. She received her B.Sc. (First Class honours), M.Sc. (with Distinction) and PhD degrees in Computer Science from Covenant University, Ota, Nigeria. She is a member of the Bioinformatics research cluster of Covenant University.

Dr Oyelade, Olanrewaju Jelili, received his Bachelor Degree in Computer Science with Mathematics (Combined Honour) and M.Sc. in Computer Science from Obafemi Awolowo University, Ile-Ife, Nigeria and PhD in Covenant University, Ota, Nigeria. He is an Associate Professor of Computer Science in the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. Dr Oyelade is a group leader of the Covenant University Bioinformatics Research Cluster (CUBRe) in the Department of Computer and Information Sciences.

Submitted: 18 January 2021; **Received (in revised form):** 4 March 2021

Introduction

Experimental approaches have been a reliable method to identify essential genes due to the extensive experimental studies on model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* [1]. However, these methods are complex, costly, labor and time-intensive. Consequently, computational approaches were deployed to complement the experimental techniques to minimize the resources required for essentiality assays. The challenge with applying the computational methods is that the quality of prediction can either enhance or burden essential genes' identification task.

Single gene deletion, antisense RNA, transposon mutagenesis and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are widely used experimental methods to identify gene essentiality. Gene deletion experiments delete a gene or pair of genes to observe the phenotype changes. This procedure must be performed thousands of times in a genome-wide study and requires extensive genome annotation. A major drawback of antisense RNA is that it is limited to the genes for which an adequate expression of the inhibitory RNA can be obtained in the organism under study. Transposon mutagenesis is the most widely used method; some of the complexities associated with this method include missing low-abundance transcripts, low resolution in locating insertion sites and narrow ranges in counting probe density [2]. CRISPR is a state-of-the-art method that provides simplicity and efficiency that can be applied directly to embryos. However, it often causes target alleles to carry additional modifications such as deletions, partial or multiple integration of the targeting vector and even duplication [3–5]. Some model organisms that were extensively studied include *Escherichia coli, Saccharomyces cerevisiae, Drosophila melanogaster, Mus musculus, Pseudomonas and Bacillus subtilis*. Their essential and nonessential gene sets have become models for poorly or understudied organisms. Due to the complexities and drawbacks of the *in vitro* approach, the computational techniques were developed to predict gene essentiality [6–8]) with the approach gaining huge popularity in recent years [9].

There are three major computational approaches available for gene essentiality prediction from peer-reviewed publications: homology mapping, constraint-based and machine learning approaches. The computational prediction method is useful when the organism is either unculturable, such as *Pneumocystis carinii*, or difficult to perform gene disruption, such as *Aspergillus fumigatus* [10]. To limit the number of failed experiments and reduce assays' cost, computational methods were used to guide candidates' choices for specific research. Examples include the use of machine learning (ML) to identify druggable and morbid genes in humans, thereby overcoming the need to perform linkage and mutation analyses to identify candidate gene(s) involved in a particular hereditary disorder [11, 12]. Also, ML was used to reveal putative Alzheimer's disease genes [13]. Most experimental methods are limited in exploring the target genome regions as cell lines are less complex compared to the whole genome [13]. ML enables the analysis of a biological or medical condition in a systemic approach since a complete genome's properties can be represented numerically. The summary of gene essentiality prediction approaches is shown in Table 1.

Homology mapping approach

Homology is the similarity in structure, physiology or development of different organisms based upon their descent from a common evolutionary ancestor [14]. It arises from duplicated genes within an organism (paralogs) or related genes in two or more different organisms (orthologs), which is a speciation product. A homology model is a form of comparative genomics (the study of differences and relationships between organisms) [15]. Homology mapping is the earliest computational approach used to determine essential genes [16]. This homology mapping requires a comparison between sequences of two organisms (a model and a target) to determine their sequence similarity based on a defined percentage identity threshold (*e*-value). If a gene sequence from a target organism shows high similarity to a sequence of an essential gene from a model organism, that gene is labeled to be essential. It premised on the biological theory that states that 'structure determines function and vice versa'.

The comparative genomic analysis includes the use of homology properties such as gene-duplication data and phyletic gene age (a measure of the most recent common ancestor) to predict essential genes. This approach has been used to predict essential genes in bacterial species such as *Mycoplasma* [17], *Liberibacter* [18], also in *P. falciparum* [19] and *Brucella spp*. [20]. This method relies only on genomic features to predict essential genes. Still, it is restricted to conserved orthologs between different species, which often make up only a small percentage of the genomes [21]. A major challenge is the significant impact of evolutionary distance on the outcome of comparative genomic analysis where an essential gene in a model organism might have its conserved ortholog to be nonessential in the target organism [22, 23]. Although essential genes tend to be highly evolutionary conserved, especially in bacteria, the conserved genes across species are not always essential [2]; this makes the homology approach less effective in essentiality prediction.

Basic Local Alignment Search Tool (BLAST) [24, 25] is one of the earliest and major tools used to perform a comparative analysis of sequences. There are five variants and several wrappers (scripts that run BLAST in a specialized way) of BLAST. The variants are BLASTN, BLASTP, BLASTX, TBLASTN and TBLASTX [25]. Some of the wrappers include PSI-BLAST, PHI-BLAST, MegaBLAST, BLASTZ, XBLAST, MPBLAST, HT-BLAST and GENE-BLAST. BLAST-like alignment tool (BLAT) is similar to BLAST with several advantages. Some of the advantages are alignment speed and the ability to submit a long list of simultaneous queries among others.

Constraint-based approach

Constraint-based methods use genome-scale metabolic networks to elucidate the biology of metabolic pathways within an organism. The metabolic network that is reconstructed based on genomic sequencing and annotations uses a constraint-based modeling technique to study the structure and function of the network's component as well as their interaction [26]. The properties of the metabolic network can be analyzed using constraint-based methods such as flux balance analysis (FBA), which predicts the fluxes of metabolites at a steady state by applying mass balance constraints to a stoichiometric model [27–30]. The concept of predicting essential genes using FBA is to simulate the knockout of a gene and evaluate the effect or impact on the network [31]. The use of FBA is better suited for studying *conditionally* essential genes because a condition can be represented as an objective function, and the significance of a gene can be determined by *in silico* deletion of the gene. The lethality is determined if there is an optimal production of predefined biosynthetic precursors. Conditionally essential genes are genes that are only essential in a given context. An example is the organism's immune response condition; genes responsible for an immune response might not be essential if

Table 1. Summary of gene essentiality prediction approaches

Attributes	Experimental	Homology mapping	Constraint-based approach	Machine learning
Description	Deactivates a gene to observe its phenotypic effect	Uses a reference model organism to classify the genes of a target organism based on the sequence similarity	Provides mathematical representation of biochemical, genetic and genomic knowledge to explain metabolic physiology	Uses a model organism to train a classifier and to classify genes of target organisms
Advantages	High accuracy in the identification of essential genes	The genomic sequence of an organism with >70% identity is sufficient to predict the essentiality of genes	Suitable for prediction of conditionally essential genes where some genes are essential in a given condition and not essential in another condition [2]	Diverse categories of features are integrated that enables it to use model organism to predict essential genes in understudied organisms
Disadvantages	Complex, costly and time-consuming [48] It is difficult to culture many of the microorganisms [51]	The approach is limited to conserved orthologs between species, which is often a small proportion of the target genome [52]. Although essential proteins tend to be conserved, there also exist large conserved nonessential proteins and essential proteins without orthologs in reference organisms [2]	Flux balance analysis, a constraint-based approach, cannot predict non-metabolic genes [2]. It requires clear definitions of nutrition availability and biomass production under precisely stated environmental conditions [53]	Available biological network data from both experimental and computational studies are incomplete and contain many false positives and false negatives, which impact the correctness of discovering essential proteins [54] . Unable to predict conditional essential genes [2, 55]
Applied organism	Bacteria and archaea [56] *Saccharomyces cerevisiae* [57]; *Schizosaccharomyces pombe* [58]; *Arabidopsis thaliana* [59]; *Mus musculus* [60] and *Homo sapiens* [61, 62]	Bacteria [63], Mycoplasma [17], Liberibacter [18], *Plasmodium falciparum* [19], and Brucella spp. [20]	Renal cell carcinoma metabolism in *Homo sapiens* [64, 65]	Salmonella typhimurium [66]; *S.cerevisiae, E.coli,* and *fungi* [67] *Homo sapiens* [68]; *Drosophila melanogaster* [69]
Scope of use	Suitable for all organisms except unculturable organisms such as *Pneumocystis carinii*, or organisms limited in genetic tractability, such as *Aspergillus fumigatus* [15].	Suitable for model organisms and organisms that are evolutionarily close to model organisms due to observed variations in gene regulations, posttranslational protein modification, divergence in cellular pathways, and redundancies in processes of distantly related species [58].	Limited to model organisms due to requirements biomass objective function, which can only be obtained from experimental studies, thus difficult to determine for non-model organisms [70].	The lack of available experimental data in most genomes limits it to model organisms and organisms that are evolutionarily close to model organisms [22].

there is no disease condition in the organism. However, they become essential when the organism is in a diseased state [32].

FBA has obvious limitations; firstly, it could only predict a metabolic gene's essentiality [2]. Secondly, unlike its ability to be integrated with modal analyses at a steady-state, FBA requires enzyme kinetic data to evaluate activities of genome-scale metabolic reactions under transient dynamic states [33]. Thirdly, upon genetic perturbation of the metabolic network, FBA fails to directly predict immediate suboptimal flux states and metabolite concentration because the organism readjusts fluxes, expressions of enzymes and various regulatory mechanisms to cushion the impact of the perturbation [9]. Lastly, FBA often requires enzyme reactions to fill gaps in the metabolic model because FBA sometimes disagrees with experimental data. It depends on the empirical models, and in some cases, parameter prediction is challenging [30, 34]. Some of these FBA limitations have been resolved through the development of FBA variants such as dynamic FBA [33], Regulatory on/off minimization (ROOM) [35], Minimization of Metabolic Adjustment (MOMA) [36] and FastMM [37].

Machine learning approach

The ability of a computer system to use statistical techniques to 'learn' and 'improve' with data to predict outcomes without being explicitly programmed accurately is known as machine learning [38]. This approach involves constructing and training one or more classifiers with training data from model organisms

composed of features of known essential genes and nonessential genes. The trained classifier is then applied to predict the essentiality of genes in the target organism. For instance, [39] generated fractal features from the genomic sequence of 27 different bacteria species and applied them to five classifiers to predict essential genes. It can be inferred that making accurate predictions requires 'good' data and an efficient machine learning technique. Supervised, semi-supervised, unsupervised and reinforcement learning are common machine learning techniques in use [40–42]. However, gene essentiality prediction is usually modeled as a classification problem, which is supervised learning.

Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. The deep learning concept is being implemented by Deep Neural Networks. There are a thousand types of deep neural network (DNN) architecture among which the following five are the state of the art: Convolutional neural networks (CNN), Recurrent neural networks (RNN), Deep Belief Networks (DBN), Variational Autoencoder and Generative Adversarial Networks (GANs). The choice of the type of deep neural network to be applied to a problem depends on two major factors: (i) the nature of the problem to be solved (voice recognition, image classification, sequence prediction, etc.), and (ii) the nature of the dataset (the type of data and how they are represented, either tabular, time series, etc.). Deep learning methods have two major drawbacks: high computational cost in the training phase and overfitting problem [43]. Deep learning techniques have been applied to many areas in bioinformatics [44]. Recently, deep learning has been used to predict essential genes [45, 46] and promising results were reported. The two major drawbacks of deep learning application to gene essentiality prediction are (i) deep neural networks require big data for training to outperform conventional ML algorithms, and (ii) complexity of the hyper-parameter tuning in deep learning models. A simple illustration of the process flow of collecting raw heterogeneous data from different sources to generate relevant features used to train a classifier and subsequently make predictions is shown in Figure 1. Data mining tools and machine learning algorithms have been used for classification. Open-source tools such as RapidMiner [47], WEKA [48], R [49] and Orange [50] provide rich functionality for data analysis and visualization.

Factors affecting the predictive performance of machine learning models for gene essentiality prediction

This review identifies some of the factors that affect machine learning models' predictive performance for essentiality prediction. These factors include (i) quality and predictive power of selected features, (ii) relatedness of training and target data and (iii) choice of the machine learning algorithm. These factors are further discussed in the following subsections.

Choice and quality of selected features

Data collection sources

Data collection and integration are the first steps in data mining or data analysis. There exist several sources of primary and secondary data for gene essentiality prediction problem, including GenBank [71] and Biomart [72] for primary sequence data, Gene Expression Omnibus (GEO) Database [73] for gene expression data, STRING Database [74, 75], BioGRID [76] for Protein interaction data, Kyoto Encyclopedia of Genomes and Genes (KEGG) [77] and BioCyc [78] for metabolic pathway data. Some of the secondary databases include Database of Essential Genes (DEG) [79], Comprehensive Microbial Resource (CMR) database at http://cmr.jcvi.org, Online Gene Essentiality database (OGEE) [80], see [2] for a comprehensive review on data sources of essential genes. With the abundance of database resources for constructing essentiality prediction models, most of them contain a significant incompleteness and error [81]. For instance, pathway databases have gaps that limit the simulation of flux distribution [82], protein interaction databases contain false positive interactions [83], among other limitations. A proven way to overcome this drawback is to use the intersection of the output from multiple sources, which reduces the degree of error in the analysis.

Feature generation

Feature generation is the process of transforming raw, unstructured data into a set of features that describes and represents the diverse attributes of the input data, often for statistical analysis or classification purposes. This process is performed after data collection and integration. In gene essentiality prediction, the input data are a set of genes or protein sequences transformed into numerical representations (features) and passed to a classifier that is expected to classify as either essential or nonessential. This set of features can be broadly categorized as intrinsic and extrinsic features, depicted in Figure 2. We define intrinsic features as features that can be directly derived from gene and protein sequences without association or comparison with another sequence; examples include gene sequence, protein sequence and Codon usage features. Features are extrinsic if they are computed from the sequence's interaction with another sequence or its environment. Examples are localization, which estimates the probability of a gene to reside in a particular compartment within the cell; topology, which computes the degree of interaction among genes or proteins; and ontology features, which encode the underrepresentation or overrepresentation of a specific gene ontology (GO) term in a given gene set (Figure 2).

Characteristics intrinsic to a gene sequence such as DNA composition, protein composition and codon usage have been used as predictors for essentiality [69, 84]. The importance of selected features in the genomic and transcriptomic categories is provided in Table 2. Extrinsic features describe a gene's essentiality from the perspective of gene expression level, functional importance and regulation complexity. Some of the features in this category include gene expression level, overrepresentation in a cellular component, molecular function, biological pathways, domain enrichment, etc. The importance of these features is described in Table 3a.

Furthermore, a topology-based feature set, a subclass of the extrinsic features, provides information on genes and gene product interaction; an example is a protein–protein interaction. The biological functions are achieved through interactions of genes and gene products, which form biological networks. Several features can be derived from the network obtained from these interactions by representing the network as a graph $G(V, E)$, where V is a set of nodes that represent genes, proteins or other components and E is a set of (directed/indirect) edges that represent their interactions. Some of the topology-based features are described in Table 3b.

Plaimas *et al.* [66] derived several features based on Lemke's (2004) definition of damage. Some of the derived features include the number of damaged reactions (NDR), number of damaged compounds (NDC), number of damaged choke point compounds (NDCC) and number of damaged chokepoint reactions (NDCR). Previous studies have shown that topology features are good

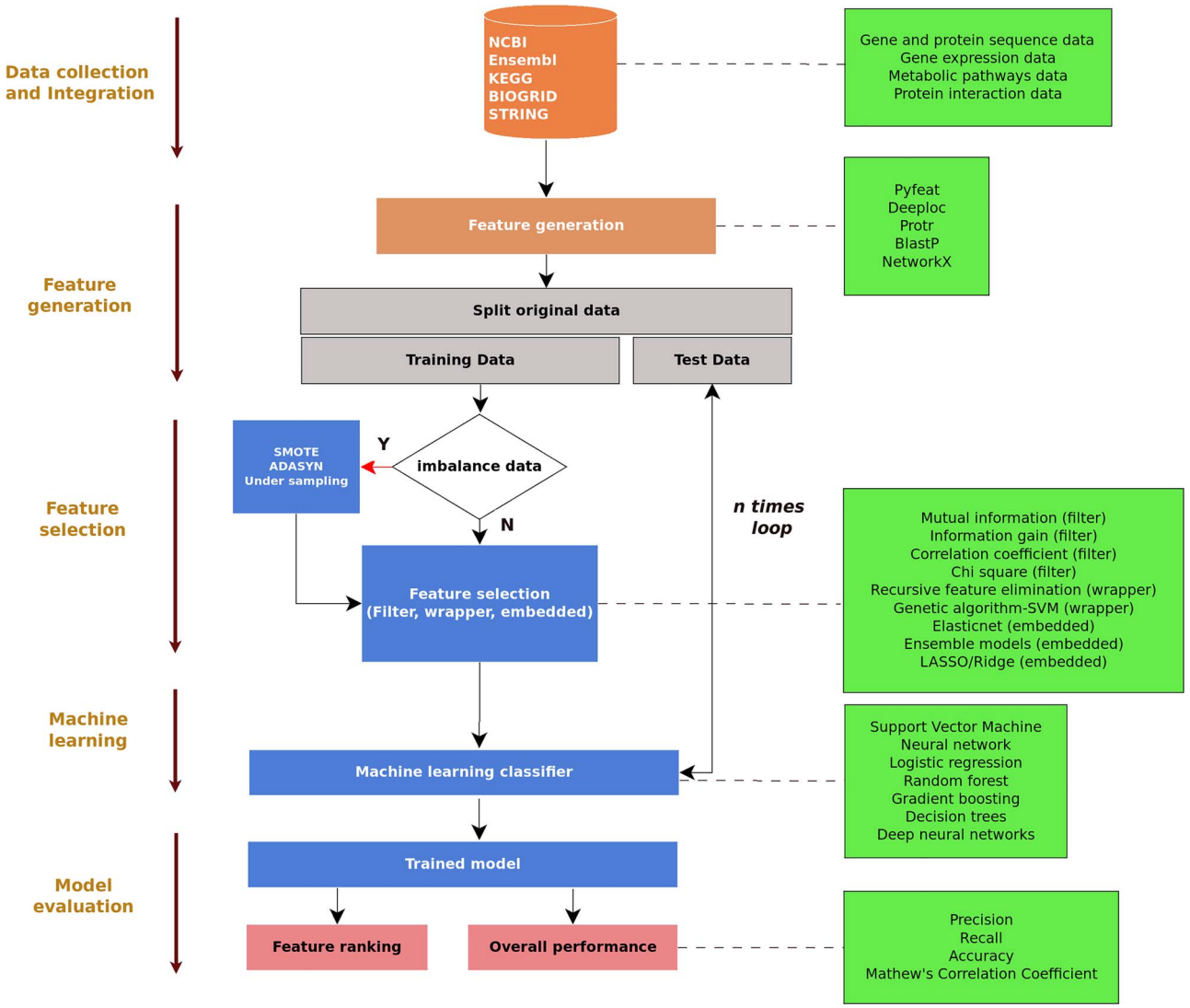

Figure 1. Simple illustration of the application of machine learning to predict essential genes.

predictors of gene essentiality [66, 112, 113]. However, these studies have not established the important categories of features that have the most discriminating power in machine learning essentiality predictions.

Most feature generation tools are freeware and easy to use. For example, *protr* [114] and *rDNAse* [115] are *R* packages for generating protein and DNA features, respectively. By simply supplying the sequences to the program in *R*, the features will be automatically generated as output. A similar procedure is applicable for *Deeploc* [116], a Linux-based tool for generating protein sub-localization features, *PSI-BLAST* [24], a stand-alone tool for generating homology features, among others. Some widely used tools for feature generation are described in Table 4. However, there are other features such as ontology features (e.g. gene ontology, KEGG orthology), topology features (e.g. number of damaged compounds [66]) that do not have standard tools to generate them. During the implementation of these features by researchers, there is a tendency for semantic errors to be introduced in the analysis, thereby affecting the analysis's eventual outcome. In addition to the lack of standard tools to generate the features above, the genome-scale databases (STRING, BioCyc, KEGG, etc.) where the sequences are retrieved often contain incomplete and error-prone data [81], which introduces bias and error into downstream analysis.

Feature selection

Before applying machine learning techniques in data analysis, one major task that must be performed is feature selection. This feature selection is necessary to reduce the dimensionality of the data and remove features that are not relevant to the classification task or could affect the quality of results or knowledge to be mined from the data. It also reduces computational time and cost. Feature selection is the process of identifying and obtaining a subset of features from a bigger set of features to enhance a classification technique's performance. There are three methods of feature selection: these are filters, wrappers and embedded. The filter technique adopts statistical evaluation methods to sort out the relevant features from the data independent of any machine learning algorithm. The major advantage of filters is the speed in selecting features but less accurate than other methods, and examples include information gain, mutual information and correlation-based feature selection [124].

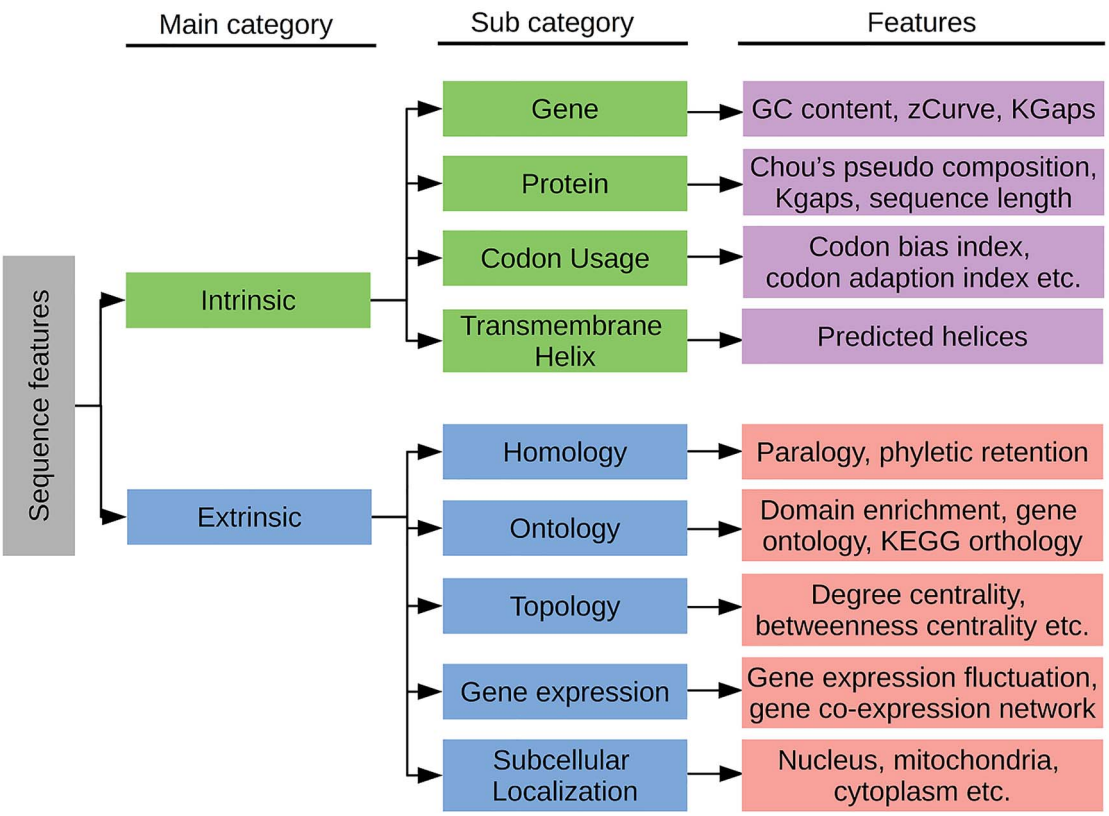

Figure 2. Categories of features for gene essentiality prediction.

Wrapper methods work by evaluating a subset of features using a machine learning algorithm that employs a search strategy to look through the space of possible feature subsets, evaluating each subset based on the quality of a given algorithm's performance. Wrapper methods are computationally intensive; examples of wrappers include Genetic Algorithm-Support Vector Machine (GA-SVM), forward selection, backward elimination and Recursive Feature Elimination (RFE) algorithm [125]. Embedded methods combine the advantages of both filter and wrapper methods by performing feature selection during the model training. Elastic Net, Ridge Regression and Random forest are examples of machine learning models that can be used for embedded feature selection [126]. The process of feature selection can significantly affect the performance of machine learning algorithms. In addition to the increase in complexity of analysis occasioned by uninformed usage of features, there is also the problem of multicollinearity (multiple features with high correlation). Failure to remove features with high correlation might affect the performance of the prediction model. The type of target organism (prokaryotes or eukaryotes) also determines the appropriate subset of essentiality prediction features, as complex organisms require more variety of features to describe essentiality [127].

In most cases, the positive samples outnumber the negative samples in the data, leading to a class imbalance problem and failure to address the problem affecting the classifier's performance. There are two major approaches used to address the class imbalance problem, namely, random oversampling and random under-sampling approaches. The oversampling approach simply generates random samples to make the minority class equals to the majority class. Synthetic Minority Oversampling Technique (SMOTE) [128] and ADAptive SYNthetic (ADASYN) [129] are state-of-the-art methods that calculate the k nearest neighbors for each sample of the minority class and randomly create multiple synthetic samples between the observation and the nearest neighbors. Depending on the number of additional samples required, ADASYN adds random small values to the synthetic samples created to add a little more variance between the minority and synthetic samples. This technique was applied in our previous study [69].

Conversely, the random under-sampling technique eliminates samples from the majority class to make the majority class equal to the minority class. This sampling technique generally leads to a reduction in data size or sample population, which often reduces machine learning power, consequently affecting model performance [130]. The random under-sampling approach is not desirable if the size of the minority class is small. Some studies have applied this method to address class imbalance [131, 132]; Nigatu *et al.* [131] even stated that the choice of a balancing approach does not influence the performance of essential gene predictions. However, we believe that using the over-sampling approach would yield better results since a higher AUC score was obtained when the over-sampling method was used, compared to under-sampling in the same study.

An extension of feature selection is feature learning, which is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data [133]. Unlike semi-structured data such as tabular data that are mathematically and computationally convenient to process, unstructured data such as images, video and sensor data are yet to have a specific feature representation

Table 2. Description of some selected intrinsic features

Subcategory	Features	Experimental importance	References
Gene sequence	*GC Content, zCurve, Kgaps, Chou's Pseudo Composition*	The stability of a DNA double helix is majorly determined by hydrogen bonds. Hence, a high number of GC pair with three hydrogen bonds will provide more hydrogen bond than AT pair with two hydrogen bonds	[85]
Protein sequence	*Chou's Pseudo Composition, KGaps,*	Characterization of protein using a matrix of amino-acid frequencies helps to deal with proteins without significant sequential homology to other proteins	[86]
Codon Usage	*Codon bias index, Codon adaptation index, Aromaticity*	These features are mostly parameters that measure optimal codon usage which determine the accurate translation of highly expressed genes, transcription control, splicing and RNA structure. Essential genes are more likely to use optimal codons. Deleterious substitution in essential genes is expected to be negligible compared to nonessential genes.	[87]
Transmembrane Helix	*Predicted helices count*	It predicts transmembrane helices and discriminates between soluble and membrane proteins	[88]

format. Generative Adversarial Networks (GANs) [134], Variational Autoencoder (VAE) [135] and AutoRegressive Networks [136] are types of deep generative models that use unsupervised learning approaches to automatically learn a set of features that best represents the data.

Feature importance in gene essentiality prediction

In this review, we sought to identify the category of features that contributes most to gene essentiality prediction performance. Gene essentiality information and features for *Caenorhabditis elegans* were assembled. *C. elegans* is one of the eukaryotic model organisms with complete sequencing and annotation of its whole genome among others such as *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*, which provided a solid foundation for structural and functional genomics explorations of the organism. Analysis of *C. elegans* essential genes in a previous study by Quin *et al.* [137] shows that they have fewer paralogs, encode proteins that are in protein interaction hubs, and are highly expressed relative to nonessential genes. These properties are similar properties to those of human disease genes. This implies that more insight about essential genes as relates to human diseases can be inferred from the outcome of this analysis.

Essentiality label for all the genes was obtained from DEG [79] and OGEE [80]. A total of 116 genes annotated as essential in both databases were selected as the positive class and other genes with ambiguous annotation were excluded from the dataset. The remaining 10 468 genes represent the negative samples. We generated commonly used 48 545 features according to five categories, namely, gene sequence (27 727 features), protein sequence (11 937 features), network topology (25 features), homology (10 features) and ontology (8846 features). Feature selection was performed to reduce the complexity of the model and a 10-fold cross-validation ML protocol was applied in which the imbalances in the class labels were corrected based on the training data. Finally, the overall performance was estimated using five performance metrics (ROC-AUC, PR-AUC, F1-score, Precision and Recall) based on the test dataset.

Light Gradient Boosting (Light GBM) ML classifier was used for the classification task. GO term feature category outperformed other categories with ROC-AUC of 0.931 and PR-AUC of 0.193. Followed by the gene ontology category is the topology category with ROC-AUC of 0.872 and PR-AUC of 0.065. Protein sequence category performs least with ROC-AUC of 0.776 and PR-AUC of 0.043 (Figure 3A). Topology (PPI) category features showed the strongest ability to distinguish essential genes from nonessential genes with an average feature importance score of approximately 2000. Homology category features ranked second to topology features also having high average feature importance scores of approximately 1000. GO terms, Gene and protein sequence features all have their feature importance score in the average of 100 as shown in Figure 3B. The superior performance of GO features is presented in Figure 3C and 3D. The poor performance based on precision, recall, F1-score and PR-AUC is due to the high imbalance in the data. The ratio of essential genes to nonessential genes is approximately 1:90.

Wang *et al.* [105] identified three criteria to consider when choosing suitable features to predict essential proteins; these are (i) ease of obtaining the feature and its availability for the target organism, (ii) high predictive capacity to identify essential proteins and (iii) the features should share minimal biological meaning. Based on the ease of obtaining features and their availability for the target organism, gene and protein sequence features rank highest because most organisms' gene and protein sequence are present in Genbank and Ensembl databases. There are bioinformatics tools for generating thousands of sequence features (Table 4). Topology features rank highest based on the high predictive capacity to identify essential genes as shown in Figure 3, with degree centrality having the highest importance of 2800. Gene ontology features can be said to share a high degree of biological meaning as this category of feature directly encodes the biological function of the genes. See Figure S1 for the feature ranking from each category.

In summary, although ontology category outperformed other feature categories based on the result of our analysis, this should be minimally applied because of its high correlation with the biological meaning of the genes. The topology category satisfies the criteria for selecting suitable features for essentiality prediction highlighted by Wang *et al.* [105]. Hence more topology-based features should be considered in essentiality prediction tasks using the machine learning approach.

Relatedness of training and investigated data

The prediction performance of machine learning classifiers is also dependent on the quality of data used to train the classifier. The quality can either be measured based on the experimental

Table 3. (a) Description of selected extrinsic features

Subcategory	Features	Experimental importance	References
Homology	*Phyletic retention*	Genes with earlier phyletic origin (older genes) are more likely to be essential and are more conserved across species than nonessential as discovered in bacteria. It is measured by the number of organisms in which an ortholog is present.	[89–91]
	Paralogy	The existence of a duplicate of a gene in a genome makes it less likely to be essential.	[87, 92]
	Orthology	Essential genes are more conserved across species than nonessential as discovered in bacteria.	[9]
	Fractals	Provides a measure of the structural complexity of the genetic sequence	[39]
Ontology	*Domain enrichment*	The existence of a functional segment of a protein sequence (domain) of a given organism in several other organisms makes it more likely to be essential.	[93, 94]
	Gene Ontology, KEGG Orthology	Functional enrichment of genes in biological process, molecular function, cellular component as well as metabolic pathways increases the likelihood of a gene to be indispensable.	[95, 96]
Protein localization	*Nucleus, cytoplasm, mitochondrion, etc.*	The location of genes in the cell is likely to determine their essentiality as essential genes are mostly located within the nuclear membrane while the nonessential are mostly found within the cytoplasm.	[7, 97, 98]
Gene expression features	*Fluctuation in gene-expression*	The fluctuation range of mRNA expression values of essential genes is often narrow while that of nonessential genes has a wide range. Essentials genes often have high expression values and are more stable.	[99, 100]
	Topology in a gene co-expression network	Similar to topology features from metabolic networks, hubs (high degree and centrality) and bottlenecks (high betweenness) are found to correlate with gene essentiality.	[101]
	Expression Profile	Genes that are not expressed under given conditions are less likely to be essential.	[102]

(b) Description of selected Network topology-based features

Feature	Description	Formula
Degree Centrality (*DC*) [103]	It describes the connectedness of a node and is the number of edges connected to a node. A gene with a high proportion of incident edges either incoming or outgoing edges is more likely to be essential.	$DC(u) = \sum_u edge(u, v)$
Betweenness Centrality (*BC*) [104]	A network attribute that quantifies the ability of the node to monitor communication between other nodes. It is defined as the average fraction of the shortest path that passes through the node [105].	$BC(u) = \sum_{i \in V} \sum_{j \in V} \frac{p(i,u,j)}{p(i,j)}, i \neq u \neq j$ $p(i, j)$ = no of the shortest paths from node i to node j, $p(i, u, j)$ = no of the shortest paths from node i to node j, which pass through node u.
Closeness Centrality (*CC*) [106]	It approximates how many edges are required to access every other reaction from a given reaction [97]. This graph attribute describes how fast a node can communicate with other nodes in a network [107].	$CC(u) = \frac{N-1}{\sum_{v \in V} dis(u,v)}$ N = no of reactions in the network
Eigenvector Centrality [108]	It is defined as the principal eigenvector of the adjacency matrix of the network and assumes that the utility of a reaction is determined by the utility of the neighboring reactions. A reaction is scored high if it is connected to high-scoring reactions [109]	$x_i = \frac{1}{\lambda} \sum_{j \in Neighbor(i)} x_j = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j$ $Neighbor(i)$ = set of neighboring reactions of reaction i, n = total number of reactions λ is a constant

(*Continued*)

Table 3. Continued

Feature	Description	Formula
Eccentricity [110]	A topological attribute that quantifies the longest distance from a specific reaction to any other reaction	$C_e(v) = \frac{n-1}{\max(d_{vi})}, i \neq v, i \in V$
Subgraph Centrality	Subgraph centrality of a node describes the number of subgraphs the node participates in.	$SC(u) = \sum_{l=0}^{\infty} \frac{u_l(u)}{l!}$ $ul(u)$ = no of closed loops of length l that starts and ends at node u.
Clustering coefficient (*CCo*) [111]	A non-supervised approach is to infer the features of an object from its neighbor	$SC(u) = \sum_{l=0}^{\infty} \frac{u_l(u)}{l!} e_i$ = no of edges connecting the adjacent nodes of node i, K_i = degree of node i.
Damage [66]	Damage is a quantitative attribute that accounts for the impact of a knocked-out reaction on a network. Given a reaction u, damage quantifies the number of reactions affected by knocking out u.	$Damage(u) = N - n$ (6) n = no of nodes in the largest cluster of the subgraph that is obtained from the network after deleting node u.

Table 4. Selected tools for generating the numerical representation of a sequence

Name	Category	Description	Reference
PyFEAT	DNA, RNA and Protein sequence	Computes the frequency distributions of various permutations of the base nucleotides/amino acids in the sequence. Similar to *rDNAse* and *Protr*	[117]
rDNAse	DNA sequence	Calculates nucleic acid composition and autocorrelation attribute of a DNA sequence	[115]
Protr	Protein sequence	Calculates state-of-the-art protein sequence descriptors such as amino acid composition and autocorrelation	[114]
CodonW	DNA and Protein sequence	Computes codon usage and correspondence analysis within a sequence	[118]
NetworkX	Network topology	Generate topology features from interaction data in Python	[119]
TopNet-like Yale Network Analyzer	Network topology	A Web system for managing, comparing and mining multiple networks	[120]
WCGNA package	Network topology	Generate topology features from gene expression data in R	[121]
PSI-BLAST	Evolution	Estimates the identity and similarity of a sequence	[24]
BUSCA	Localization	A Web system for predicting the subcellular localization	[122]
TmHMM Server	TmHMM	An online system for predicting transmembrane helices in proteins	[88]
ProPAS	Physico-chemical	Calculates the Isoelectric point (pI), Mass weight (MW) and Hydrophobicity (Hy) properties of protein sequences	[123]
Deeploc	Localization	A stand-alone application for predicting subcellular localization of proteins	[116]

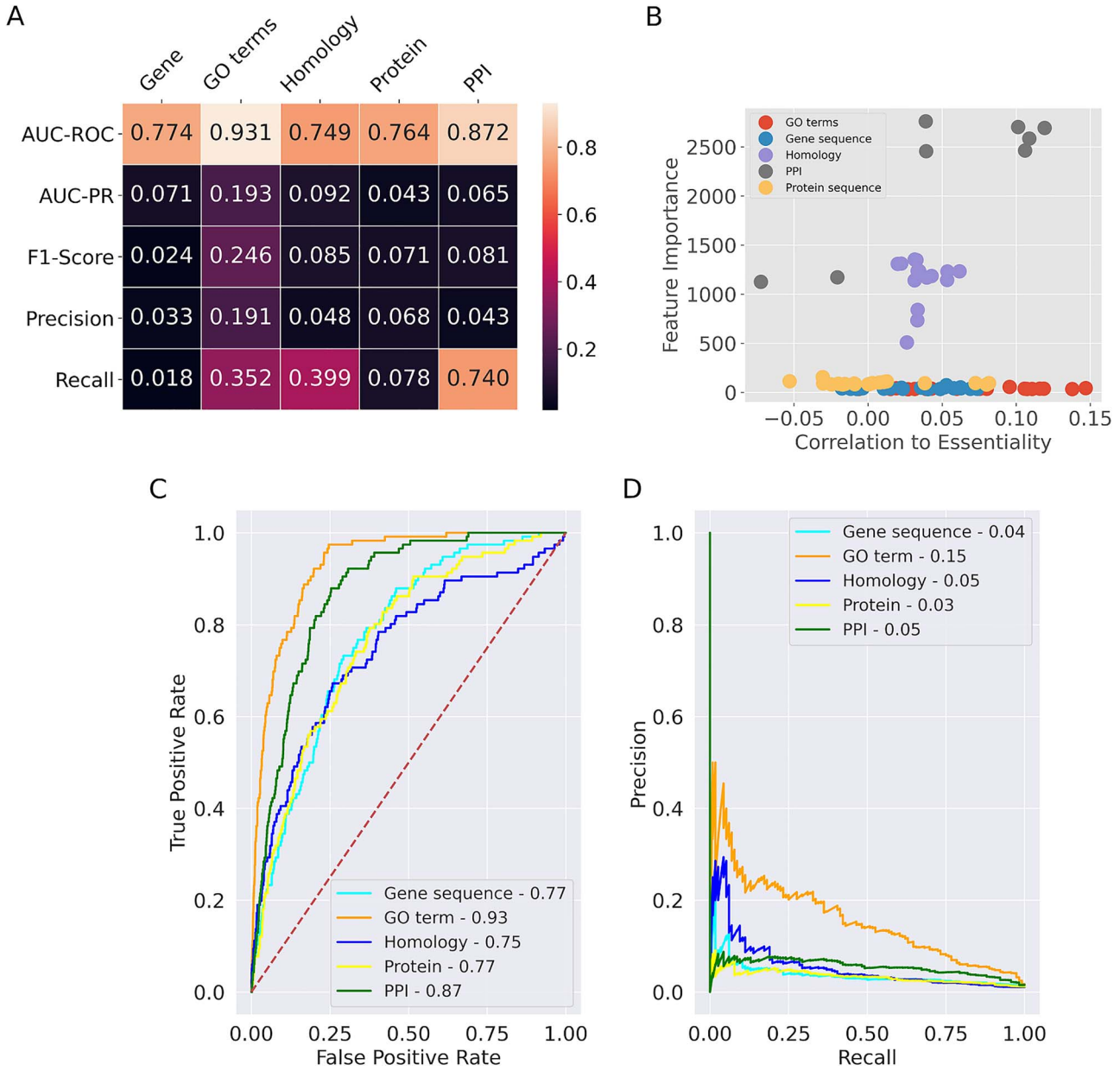

Figure 3. The performance evaluation of five categories of features generated from *C. elegans*, using 10-fold CV of light GBM classifier. (A) Heatmap shows GO terms feature category having superior scores except in recall where topology (PPI) features performs better. (B) Feature importance of the feature categories shows three interesting clusters with PPI features having highest importance in the top cluster, homology features occupy the middle cluster while GO terms, gene sequence and protein categories cluster at the bottom. (C) Shows GO terms category has superior ROC-AUC. (D) PR-AUC curve shows performance difference between all categories of features with GO terms category having superior performance.

process' accuracy that produces the data or the data's closeness to the target domain. Cheng *et al.* [130] stated that essential genes obtained from genome-wide gene deletion experiments produced superior quality datasets than essential genes identified through transposon mutagenesis, RNA interference (RNAi) and other methods. It is a general principle in machine learning; the bigger the training data's size, the better the predictive model's performance. This was also validated by Cheng *et al.* [130] that varied the input data's size and concluded that there was a significant improvement in the machine learning model's robustness and predictive accuracy. However, the data quality referred to in this study is the closeness or correlation of the input data to the target domain or the question the study sought to answer. For instance, the data collected on breast cancer will perform optimally when used to investigate breast cancer than when used to investigate prostate cancer.

The prediction of a gene's essentiality with a machine learning approach, the classifier can be trained by learning the characteristics from known essential genes of the target organism (intraspecies) or transferring essential gene annotations from a closely related model organism (cross-organism) [10]. However, for an understudied microbe, each approach has its potential limitations; the intraspecies approach is constricted by the often-small number of known essential genes. The cross-organism approach is limited by the availability of model organisms closely related to the understudied organism based on

evolutionary distance. Deng *et al.* [10] investigated three approaches to predict essential genes based on the number of known essential genes by studying four bacteria organisms to validate the impact of relatedness on prediction performance. Two prokaryotes (*Escherichia coli K-12* and *Acinetobacter baylyi* ADP1) and two eukaryotes (*Saccharomyces cerevisiae* S288c and *Neurospora crassa* OR74A) were evaluated individually and in pairs. The first (*intra*) approach which requires learning from the known essential genes in the target organism will be suitable if the number of known essential genes is at least 2% and 4% of the total genes in prokaryotes and eukaryotes, respectively. The second (*cross*) approach which involves transferring essential gene annotations from a related model organism is suitable when the number of known essential genes is less than 2% (4% in eukaryotes) of the total genes and there is a closely related organism. The third approach (hybrid) combines both approaches and outperforms both of them when applied to an understudied organism. Peng *et al.* [2] also stated there is improved prediction performance if the organism under study belongs to the same phylogenetic lineage with the reference species. Hence, it can be inferred that relatedness of data can significantly improve the performance of predictive models.

Application of machine learning in predicting essential genes

Gene essentiality prediction is naturally a classification problem because a set of label samples representing all possible classes within the population is used to classify unlabeled samples. Therefore, several supervised learning approaches have been used to predict gene or protein essentiality. There is no single ML algorithm that performs best in all domains or given different problems or data types [38]. Not all machine learning algorithms are suitable for essentiality prediction as a wrong choice will produce a poor prediction, meaning that the quality of prediction is also dependent on the choice of machine learning algorithm used. A summary of some selected studies that applied machine learning techniques to predict essential genes (Table 5) shows that the performance of each technique is due to factors such as quality of features and the type of algorithm used to train the classifier. Some studies focus on a single organism to train and test the classifier (intra-organism). However, some combined data from different but similar organisms (cross-organism) to train and test the classifier.

The application of a broad collection of intrinsic and extrinsic features has been shown to improve the performance of gene essentiality classifiers [69]. Also, the power of deep neural networks is yet to be fully exploited in essentiality prediction problems as recent studies that applied deep neural networks to predict essential genes have reported superior performance and accuracy compared to conventional classifiers like SVM, Random forest, Decision tree, Logistic regression among others [45, 46].

Prediction evaluation

In determining essential genes using the computational approach, it is pertinent to validate the model's predictions and evaluate its performance. This is because poor prediction accuracy will defeat the basic purpose of applying computational approaches to complement experimental approaches. A model is evaluated based on its performance on unseen data and not the training data. As a standard practice to obtain reliable estimates of a machine learning model's performance, *k*-fold cross-validation is among the reliable methods available,

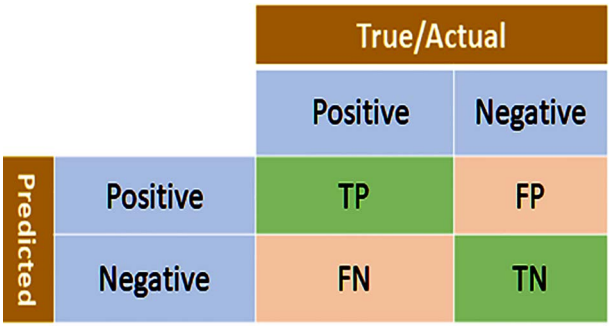

Figure 4. Confusion matrix (TP = True positive, FP = False positive, TN = True negative, FN = False negative)

particularly for small datasets. *K*-fold cross-validation technique operates by randomly dividing the training data into k groups, using the *k*-1 groups to construct the model, and using the *k*th group to evaluate the model. It stores the performance of the model and repeats the process for the remaining groups.

One of the methods used to evaluate the model's performance is by comparing the predicted values to the real values for regression problems. This can be achieved by estimating either Mean Absolute Error (MAE) (see Equ.1), Mean Squared Error (MSE) (see Equ.2) or Normalized Mean Squared Error (NMSE) (see Equ.3) [142].

$$MAE = mean\left(absolute\left(Value_{pred} - Value_{real}\right)\right) \quad (1)$$

$$MSE = mean\left(square\left(Value_{pred} - Value_{real}\right)\right) \quad (2)$$

$$NMSE = \frac{mean\left(square\left(Value_{pred} - Value_{real}\right)\right)}{mean\left(square\left(Value_{real} - Value_{realpred}\right)\right)} \quad (3)$$

The use of MSE has the disadvantage of not being measured in the same unit as the target variable, making it difficult to interpret from the user's perspective. There is no benchmark to determine if a prediction is good or not when MAE is used as the evaluation metric. Notwithstanding, NMSE is a better statistic that calculates the ratio between the model's output and a benchmark value. NMSE values range from 0 to 1 where values close to 0 indicate good performance and values close to 1 indicate that the model performs worse and not predictive [142]. The set of standard metrics used to evaluate a binary classification model's performance is depicted in Table 6. To compute the evaluation metrics, basic parameters such as True positives (TP), False positives (FP), True Negatives (TN) and False negatives (FN) must be calculated (Figure 4). A prediction is TP if correctly predicted as Positive; FP if wrongly predicted as Positive. TN is correctly predicted as Negative, and FN is wrongly predicted as Negative.

Evaluation of predictive models for binary classification using precision, recall, and by extension, F1-score considers only the positive class as the class of interest while neglecting the negative class. They use only three of the confusion matrix values: TP, FP and FN, while the 4th value, TN, is not used in these metrics. This means that the value of TN is not considered in the model evaluation. Although accuracy uses all the confusion matrix values, it produces a misleading measurement if the samples in one class are more than the other class. Gene essentiality prediction focuses only on the positive class (essential genes); therefore, these metrics will provide a good performance evaluation. However, if both positive and negative classes are of interest, then Matthews Correlation Coefficient (MCC) will be a better measure. MCC considers all

Table 5. Performance of some selected studies that used machine learning techniques to predict gene essentiality

ML algorithms	features	Type of prediction	Purpose of the study	Performance
NN, SVM [6]	DC and gene expression	Intra-organism	Essential protein prediction in *E. coli* and *S. cerevisiae*	ROC-AUC ranges from 0.69 to 0.89
WKNN, SVM, Ensemble [138]	DC and Sequence-related	Intra-organism	Essential gene prediction in *S. cerevisiae*	Recall ranges from 0.73 to 0.81
NB [7]	DC and Sequence-related	Intra-organism	Essential gene prediction in *E. coli* and *S. cerevisiae*	ROC-AUC ranges from 0.6984 to 0.7004
C4.5 decision tree [139]	DC	Intra-organism	Essential gene prediction in *E. coli*	F-measure ranges from 0.797 to 0.834
SVM [111]	Centrality and Sequence-related	Intra-organism	Essential gene prediction in *E. coli* and *S. cerevisiae*	Precision ranges from 0.72 to 0.83
Single and Ensemble Decision tree [97]	Centrality, localization and gene ontology terms	Intra-organism	Essential gene prediction in *S. cerevisiae*	ROC-AUC within the range 0.667 to 0.808
SVM [66]	Centrality	Cross-organism	Essential gene prediction in distantly related bacteria	ROC-AUC within the range of 0.75 to 0.81
Ensemble [10]	Centrality, gene expression, and Sequence-related	Cross-organism	Essential gene prediction in distantly related bacteria	ROC-AUC scores within 0.69 and 0.89
FWM (NB, LR, genetic algorithm) [127]	Centrality, gene expression, and Sequence-related	Cross-organism	Essential gene prediction in bacteria species	ROC-AUC ranges from 0.77 to 0.95
NB, LR, C4.5 decision tree and CN2 rule [15]	Gene expression and Sequence-related	Cross-organism	Essential gene prediction in eukaryotic fungal species to identify potential drug target	ROC-AUC scores between 0.69 and 0.89
NB [130]	Centrality, gene expression and Sequence-related	Cross-organism	Essential gene prediction in bacteria species	ROC-AUC scores range from 0.781 to 0.941
SVM [96]	Gene ontology terms and KEGG pathways	Intra-organism	Essential gene prediction in human leukemia cell line	MCC score of 0.951
SVM [140]	Centrality and Sequence-related	Cross-organism	Essential gene prediction in distantly related bacteria	ROC-AUC scores of 0.857 and precision of 0.335
Ensemble [141]	Centrality, gene expression and Sequence-related	Intra-organism	Prediction of essential protein in *S. cerevisiae* using a unique network centrality feature	Precision scores between 0.651 and 0.862
Deep Neural Network [45]	Topology features, gene expression profiles, localization	Intra-organism	Essential protein prediction in *S. cerevisiae*	ROC-AUC scores between 0.831 and 0.841
Deep Neural Network [46]	Sequence features	Cross-organism	Essential gene prediction in bacteria species	ROC-AUC scores between 0.838 and 0.842
Ensemble [69]	Sequence, Topology, Homology, Ontology, Localization	Intra-organism	Prediction of essential genes in *D. melanogaster*	ROC-AUC score of 0.922

Abbreviations: NN, neural network; DC, degree centrality; ROC-AUC, area under the receiver operating characteristic curve; WKNN, weighted k-nearest-neighbor; SVM, support vector machine; NB, Naive Bayes; FWM, feature-based weighted Naïve Bayes model; PIN, protein–protein interaction network; LR, logistic regression; MCC, Matthews correlation coefficient.

four values in the confusion matrix, and a high value (close to 1) means that both classes are predicted well, even if one class is disproportionately (under or over) represented.

Area Under Receiver Operating Characteristic (ROC-AUC) curve graphically represents the trade-off between true positive rate (sensitivity) and false positive rate (1—specificity) of a given model at different thresholds. It is used to select optimal binary classifiers independently from class distribution, and its scores range from 0 to 1. Area Under Precision-Recall curves (PR-AUC) graphically represents the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. The Precision-Recall plot is more informative than the ROC-AUC plot when evaluating binary classifiers on imbalanced datasets [146].

Challenges in machine learning approach for future research

The computational approaches have significantly bridged the gaps identified in the experimental approaches to predicting essential genes. However, there still exist some areas that need to be improved upon. First, information from model organisms used to train the classifiers is often incomplete and contains false positives and false negatives due to experimental errors

Table 6. Standard evaluation metrics for binary classification

Metric	Description	Formula
Accuracy	It measures the degree of correctness of a model with respect to both positive and negative classes.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision or Positive Predictive Value (PPV)	It measures the ratio of true positives with regards to all positives predicted by the model [143].	$\frac{TP}{TP+FP}$
Sensitivity or Recall	Measures the proportion of actual positives that are correctly identified as such. Also known as True positive rate [143].	$\frac{TP}{TP+FN}$
Specificity	Measures the proportion of actual negatives that are correctly identified as such. Also known as True negative rate.	$\frac{TN}{TN+FP}$
F1-score	It is the harmonic mean of recall and precision	$F_1 = (\frac{recall^{-1}+precision^{-1}}{2})^{-1} = 2 \cdot \frac{precision \cdot recall}{precision+recall}$
Matthews correlation coefficient (MCC)	It is a balanced measurement even if the sizes of positive and negative samples have a great difference. The coefficient ranges between +1 and − 1. 1 represents a perfect prediction, 0 is better than random prediction and − 1 indicates total disagreement between prediction and observation [144].	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Cohen's Kappa	It measures the degree of agreement between the output of two models. If $\kappa = 1$ then the models are in perfect agreement and $\kappa = 0$ means there is no agreement between models [145]	$\kappa \equiv \frac{p_0 - p_e}{1 - p_e}$ p_0 is the relative observed agreement among models, p_e is the hypothetical probability of chance agreement.

and consequently, affects the classifier [84, 105]. Recent publications have revealed that studies that assemble their class labels from multiple sources (databases or studies) produce superior results compared to those that used a single source [45, 69]. Second, the limited availability of model organisms and evolutionary distance to the target organism is another challenge because the prediction performance deteriorates as distance increases.

Third, several studies have reported that some nonessential genes became essential when placed in different environmental conditions [105, 147, 148]. This poses a big challenge to ML techniques due to label inaccuracy and inconsistency. For instance, a study of gene essentiality experiment by Juhas *et al.* [23] reported inconsistent results of essential genes in the same organisms under similar experimental conditions. This lack of consensus makes it difficult to determine gene essentiality in model organisms, let alone in non-model or poorly researched organisms. Moreover, the prediction of essential genes is significantly affected by the upregulation of isoenzymes which occurs as a result of the longer duration required for conducting experiments [9]. The reliability of an ML approach is questionable if the class label for the training data is based on one experimental condition only. Consequently, studies that combine essentiality labels from multiple experiments for ML analysis will produce an improved performance for absolute gene essentiality prediction.

The differences in the outcome of gene essentiality studies are possibly a result of 'conditional' or contextual essentiality because the essentiality of a gene depends on its context, which might be a defined growth media or conditions, genetic context, or a particular developmental stage of a microorganism [149]. This leads to the more specialized use of ML to predict conditionally essential genes, which is the fourth challenge identified by this review. The application of ML techniques to predict conditionally essential genes has not been suitable [2] majorly because there are insufficient labels to train predictive models for diverse biological conditions in several organisms. For instance, the use of ML techniques to predict essential genes in immune response condition requires annotating class labels from the literature, which might not be sufficient to train an ML model thereby making the approach inappropriate. Hence, there is a need to develop protocols and techniques capable of predicting conditional gene essentiality using machine learning.

Cooperative machine learning (CML) [150] could answer the limitation of machine learning techniques in predicting conditionally essential genes. CML is an extension of active learning that efficiently combines human intelligence with the machine's rapid computation ability to annotate unlabeled data. The trained classifier uses the small labeled data for its training as shown in Figure 5. Therefore, given a small labeled data of a condition of interest, CML can be used to predict the unlabeled section of the data. Another interesting technique that could be applied to predict conditionally essential genes is a deep generative model such as Generative Adversarial Networks (GANs). GANs are state-of-the-art deep learning models that can model high-dimensional data, handle missing data, and can provide multimodal outputs. The generative modeling provided by GANs makes it a potentially suitable technique to address the limitation of ML in predicting conditional essential genes. Conditional GANs can be used to generate new examples for tasks that require more samples for model training, thereby making it a plausible technique for conditional essentiality prediction. See the conference report by Ian Goodfellow for details about GANs [151].

Absence of a central store for documenting experimentally identified essential sets of genes and gene products was also identified in this study as major challenges. Since the machine learning approach requires annotation of model organisms with experimentally determined gene essentiality for its construction, only a few model organisms [such as *Drosophila melanogaster* [152], *Mus musculus* [153] and *Saccharomyces cerevisiae* [154] have a dedicated database for essentiality annotation.

Conclusion

Identification of essential genes is imperative because it provides an understanding of the core structure and function of a cell, accelerates the discovery of drug targets, guides the engineering of new organisms, provides knowledge about the basic requirements for a cell, and proffers insights into the correlations between genotype and phenotype. Computational approaches have been advanced to serve as alternative

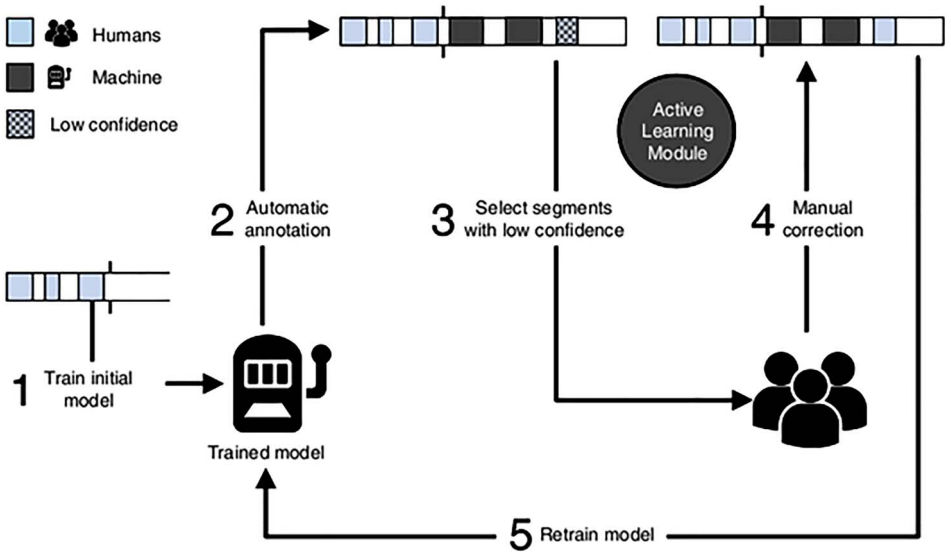

Figure 5. Conceptual framework for cooperative machine learning (1) the initial model is trained using the limited labeled data (2) the trained model is used to predict the unlabeled data (3) predictions with low confidence is selected (4) and manually annotated by the oracle (5) the initial model is retrained using the predicted added to the labeled data [155].

and complement experimental approaches. Three important factors that determine the prediction performance of machine learning methods are (i) predictive power of selected features (ii) relatedness of training and prediction dataset and (iii) suitability of machine learning algorithm. Topology features were identified to possess high discriminating power in gene essentiality prediction and less biological correlation to the gene function, thus making it a highly suitable feature category for essentiality prediction. Embedded feature selection methods are practically suitable compared to other techniques due to their ability to handle large feature sets and superior performance. Several challenges in using computational approaches to predict essential genes were highlighted with a view of further studies in them. State-of-the-art machine learning approach such as deep learning provides a brighter prospect of developing prediction models that can predict essential genes in related organisms and distantly related organisms by taking advantage of the automatic feature selection and accuracy in deep learning methods. Furthermore, cooperative machine learning and Generative Adversarial Networks could be further exploited to develop models that can perform conditional essentiality predictions.

Key Points

- The choice of features and ML technique is vital to gene essentiality prediction.
- Most data sources contain a significant measure of incompleteness and error that affects downstream ML analysis.
- Feature category such as gene ontology that has a high correlation with the biological meaning of the genes should be minimally applied.
- Cooperative machine learning technique could provide an answer to the prediction of conditionally essential genes.

Author's contribution

All authors contributed to this work. O.A. contributed to the original idea and conception. O.A. achieved the implementation of the comparative analysis. O.A. and D.A. performed data curation and Visualization. J.O., I.I. and O.A. wrote the manuscript. O.A., J.O., I.I. and D.A. were responsible for the final version's critical revision and approval.

Acknowledgments

This work was supported by the Covenant University Center for Research, Innovation, and Discovery (CUCRID), the Deutsche Forschungsgemein schaft (https://www.dfg.de/) within the project KO 3678/5-1, and the German Federal Ministry of Education and Research (BMBF) within the project Center for Sepsis Control and Care (CSCC, 01EO1002, and 01EO1502). We also thanked Prof. Konig Rainer and Dr. Thomas Beder for their useful contributions.

References

1. Hart T, Brown KR, Sircoulomb F, *et al*. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* EMBO Press 2014;**10**:733.
2. Peng C, Lin Y, Luo H, *et al*. A comprehensive overview of online resources to identify and predict bacterial essential genes. *Front Microbiol* Frontiers 2017;**8**:2331.
3. Li J, Shou J, Guo Y, *et al*. Efficient inversions and duplications of mammalian regulatory DNA elements and gene clusters by CRISPR/Cas9. *J Mol Cell Biol* Oxford University Press 2015;**7**:284–98.
4. Pavlovic G, Erbs V, Andre P, *et al*. Generation of targeted overexpressing models by CRISPR/Cas9 and need of careful validation of your knock-in line obtained by nuclease genome editing. *Transgenic Res* 2016;**25**:254–5.

5. Flora A, Welcker J. CRISPR Genome Engineering: Advantages and Limitations., Rodent Research Models 2017;**22**
6. Chen Y, Xu D. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* 2005 [cited 2018 May 17];**21**:575–81. Available from. http://www.ncbi.nlm.nih.gov/pubmed/15479713.
7. Gustafson AM, Snitkin ES, Parker SC, *et al*. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* 2006;**7**:265.
8. Seringhaus M, Paccanaro A, Borneman A, *et al*. Predicting essential genes in fungal genomes. *PCR Methods Appl*. Cold Spring Harbor Laboratory Press 2006 [cited 2018 Jul 14];**16**:1126–35. Available from. http://www.ncbi.nlm.nih.gov/pubmed/16899653.
9. Mobegi FM, van Hijum SAFT, Burghout P, *et al*. From microbial gene essentiality to novel antimicrobial drug targets. *BMC Genomics* BioMed Central 2014;**15**:958.
10. Deng J, Deng L, Su S, *et al*. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res* 2011 [cited 2018 May 17];**39**:795–807. Available from. http://www.ncbi.nlm.nih.gov/pubmed/20870748.
11. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* Oxford University Press 2009;**37**:D767–72.
12. Costa PR, Acencio ML, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics* Springer 2010;**11**:1–15.
13. Huang X, Liu H, Li X, *et al*. Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. *BMC Neurol* Springer 2018;**18**:5.
14. Panchen AL. Homology-history of a concept. *Novartis Found Symp* Wiley Online Library 1999;**225**:5–18.
15. Lu Y, Deng J, Rhodes JC, *et al*. Predicting essential genes for identifying potential drug targets in aspergillus fumigatus. *Comput Chem*. Elsevier 2014 [cited 2018 May 17];**50**:29–40. Available from. https://www.sciencedirect.com/science/article/pii/S1476927114000139.
16. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci* National Acad Sciences 1996;**93**:10268–73.
17. Liu W, Fang L, Li M, *et al*. Comparative genomics of mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. *PLoS One* Public Library of Science 2012;**7**:e35698.
18. Fagen JR, Leonard MT, McCullough CM, *et al*. Comparative genomics of cultured and uncultured strains suggests genes essential for free-living growth of Liberibacter. *PLoS One* Public Library of Science 2014;**9**:e84469.
19. Rout S, Warhurst DC, Suar M, *et al*. In silico comparative genomics analysis of plasmodium falciparum for the identification of putative essential genes and therapeutic candidates. *J Microbiol Methods* Elsevier 2015;**109**:1–8.
20. Yang X, Li Y, Zang J, *et al*. Analysis of pan-genome to identify the core genes and essential genes of Brucella spp. *Mol Genet Genomics* Springer 2016;**291**:905–12.
21. Zdobnov EM, von Mering C, Letunic I, *et al*. Paucity of genes on the drosophila X chromosome showing male-biased expression. *Science (80-)* [Internet]. American Association for the Advancement of Science 2002 [cited 2019 Oct 25];**298**:149–59. Available from. http://www.sciencemag.org/lookup/doi/10.1126/science.1077061.
22. Wei W, Ning L-W, Ye Y-N, *et al*. Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One* Public Library of Science 2013;**8**:e72343.
23. Juhas M, Eberl L, Glass JI. Essence of life: essential genes of minimal genomes. *Trends Cell Biol* Elsevier 2011;**21**:562–8.
24. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* Oxford University Press 1997;**25**:3389–402.
25. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. *Nucleic Acids Res* Oxford University Press 2006;**34**:W6–9.
26. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* [Internet]. NIH Public Access 2010 [cited 2018 May 15];**5**:93–121. Available from. http://www.ncbi.nlm.nih.gov/pubmed/20057383.
27. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* Elsevier 2003;**14**:491–6.
28. Papp B, Pal C, Hurst LD. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* Nature Publishing Group 2004;**429**:661–4.
29. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform* Oxford Univ Press 2009;**10**:435–49.
30. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* Nature Publishing Group 2010;**28**:245.
31. Basler G. Computational prediction of essential metabolic genes using constraint-based approaches. *Gene Essentiality* Springer 2015;**1279**:183–204.
32. Levashina EA. Immune responses in Anopheles gambiae. *Insect Biochem Mol Biol* 2004;**34**:673–8.
33. Mahadevan R, Edwards JS, Doyle FJ, III. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys J* Elsevier 2002;**83**:1331–40.
34. Zomorrodi AR, Suthers PF, Ranganathan S, *et al*. Mathematical optimization applications in metabolic networks. *Metab Eng* Elsevier 2012;**14**:672–86.
35. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci* National Acad Sciences 2005;**102**:7695–700.
36. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci* National Acad Sciences 2002;**99**:15112–7.
37. Li G-H, Dai S, Han F, *et al*. FastMM: an efficient toolbox for personalized constraint-based metabolic modeling. *BMC Bioinformatics BioMed Central* 2020;**21**:1–7.
38. Sakr S, Elshawi R, Ahmed AM, *et al*. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercise testing (FIT) project. BMC med inform Decis Mak. *BioMed Central* 2017;**17**:174.
39. Yu Y, Yang L, Liu Z, *et al*. Gene essentiality prediction based on fractal features and machine learning. *Mol Biosyst Royal Society of Chemistry* 2017;**13**:577–84.
40. Baştanlar Y, Özuysal M. Introduction to machine learning. *miRNomics MicroRNA Biol Comput Anal* Springer 2014;**1107**:105–28.
41. Evers B, Jastrzebski K, Heijmans JPM, *et al*. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying

essential genes. *Nat Biotechnol* Nature Publishing Group 2016;**34**:631.
42. Adamu PI, Aromolaran O. Machine learning priority rule (MLPR) for solving resource-constrained project scheduling problems. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 2, Hongkong, 2018.
43. Pasupa K, Sunhem W. A comparison between shallow and deep architecture classifiers on small dataset. In: 8th International Conference on Information Technology and Electrical Engineering (ICITEE) 5, p. 1–6, Indonesia 2016.
44. Li Y, Huang C, Ding L, *et al*. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods Elsevier* 2019;**166**:4–21.
45. Zeng M, Li M, Fei Z, *et al*. A deep learning framework for identifying essential proteins by integrating multiple types of biological information. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:296–305.
46. Hasan MA, Lonardi S. DeeplyEssential: a deep neural network for predicting essential genes in microbes. *bioRxiv Cold Spring Harbor Laboratory* 2019;607085.
47. Mierswa I, Klinkenberg R. RapidMiner Studio (9.2)[Data science, machine learning, predictive analytics]. 2018. Available online: https://rapidminer.com/ (accessed on 09 November 2020).
48. Witten IH, Frank E, Hall MA, *et al*. *Data mining fourth edition: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann Publishers Inc, 2016.
49. R Core Team *D. A Language and Environment for Statistical Computing [Internet]*. R Found. Stat. Comput. Vienna, Austria; 2020. p. https://www.R-project.org. Available from: http://www.r-project.org
50. Demšar J, Curk T, Erjavec A, *et al*. Orange: data mining toolbox in python. *J Mach Learn Res* JMLR. org 2013;**14**:2349–53.
51. Lin Y, Zhang F-Z, Xue K, *et al*. Identifying bacterial essential genes based on a feature-integrated method. *IEEE/ACM Trans Comput Biol Bioinform* IEEE 2019;**16**:1274–9.
52. Bruccoleri RE, Dougherty TJ, Davison DB. Concordance analysis of microbial genomes. *Nucleic Acids Res* Oxford University Press 1998;**26**:4482–6.
53. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol* EMBO Press 2007;**3**:119.
54. Marcotte EM, Pellegrini M, Thompson MJ, *et al*. A combined algorithm for genome-wide prediction of protein function. *Nature* Nature Publishing Group 1999;**402**:83–6.
55. Mobegi FM, Zomer A, de Jonge MI, van Hijum SAFT. Advances and perspectives in computational prediction of microbial gene essentiality. *Brief Funct Genomics Oxford University Press*; 2016;**16**:70–9.
56. Giaever G, Chu AM, Ni L, *et al*. Functional profiling of the Saccharomyces cerevisiae genome. *Nature [Internet]* 2002 [cited 2018 May 15];**418**:387–91. Available from. http://www.ncbi.nlm.nih.gov/pubmed/12140549.
57. Sarmiento F, Mrázek J, Whitman WB. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis. *Proc Natl Acad Sci* National Acad Sciences 2013;**110**:4726–31.
58. Kim D-U, Hayles J, Kim D, *et al*. Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* Nature Publishing Group 2010;**28**:617.
59. Meinke D, Muralla R, Sweeney C, *et al*. Identifying essential genes in Arabidopsis thaliana. *Trends Plant Sci* Elsevier 2008;**13**:483–91.
60. Liao B-Y, Zhang J. Mouse duplicate genes are as essential as singletons. *Trends Genet* Elsevier 2007;**23**:378–81.
61. Blomen VA, Májek P, Jae LT, *et al*. Gene essentiality and synthetic lethality in haploid human cells. Science (80-). American association for the. *Adv Sci* 2015;**350**:1092–6.
62. Wang T, Birsoy K, Hughes NW, *et al*. Identification and characterization of essential genes in the human genome. Science (80-). American association for the. *Adv Sci* 2015;**350**:1096–101.
63. Hua H-L, Zhang F-Z, Labena AA, *et al*. An approach for predicting essential genes using multiple homology mapping and machine learning algorithms. *Biomed Res Int* Hindawi 2016;**2016**:7639397.
64. Zhong J, Wang J, Peng W, *et al*. Prediction of essential proteins based on gene expression programming. BMC genomics. *BioMed Central* 2013;**14**:S7.
65. Gatto F, Miess H, Schulze A, *et al*. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. *Sci Rep* Nature Publishing Group 2015;**5**:10738.
66. Plaimas K, Eils R, König R. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol [Internet]* 2010 [cited 2018 May 3];**4**:56. Available from. http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-4-56.
67. Deng J, Tan L, Lin X, *et al*. Exploring the optimal strategy to predict essential genes in microbes. Biomolecules. *Molecular Diversity Preservation International* 2011;**2**:1–22.
68. Chen H, Zhang Z, Jiang S, *et al*. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. *Brief Bioinform* 2020;**21**:1397–410.
69. Aromolaran O, Beder T, Oswald M, *et al*. Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. *Comput Struct Biotechnol J* Elsevier 2020;**18**:612–21.
70. Yuan H, Cheung CY, Hilbers PAJ, *et al*. Flux balance analysis of plant metabolism: the effect of biomass composition and model structure on model predictions. *Front Plant Sci* Frontiers 2016;**7**:537.
71. Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* Oxford University Press 2017;**45**:D37–42.
72. Smedley D, Haider S, Ballester B, *et al*. BioMart–biological queries made easy. BMC genomics. *BioMed Central* 2009;**10**:22.
73. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* Oxford University Press 2012;**41**:D991–5.
74. Jensen LJ, Kuhn M, Stark M, *et al*. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* Oxford University Press 2008;**37**:D412–6.
75. Szklarczyk D, Gable AL, Lyon D, *et al*. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* Oxford University Press 2018;**47**:D607–13.
76. Oughtred R, Stark C, Breitkreutz BJ, *et al*. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* Oxford University Press 2019;**47**:D529–41.
77. Kanehisa M, Sato Y, Kawashima M, *et al*. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* Oxford University Press 2016;**44**:D457–62.

78. Caspi R, Altman T, Dreher K, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* Oxford Univ Press 2012;**40**:D742–53.
79. Luo H, Lin Y, Gao F, *et al*. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res* Oxford University Press 2014;**42**:D574–80.
80. Chen W-H, Lu G, Chen X, *et al*. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res* Oxford University Press 2017;**45**:D940–4 gkw1013.
81. Zhang F, Peng W, Yang Y, *et al*. Novel method for identifying essential genes by fusing dynamic protein–protein interactive networks. Genes (Basel). *Multidisciplinary Digital Publishing Institute* 2019;**10**:31.
82. Wang H, Marcišauskas S, Sánchez BJ, *et al*. RAVEN 2.0: a versatile platform for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLoS Comput Biol* Cold Spring Harbor Laboratory 2018;**14**:e1006541.
83. Murali TM, Dyer MD, Badger D, *et al*. Network-based prediction and analysis of HIV dependency factors. *PLoS Comput Biol* Public Library of Science 2011;**7**:e1002164.
84. Campos TL, Korhonen PK, Gasser RB, *et al*. An evaluation of machine learning approaches for the prediction of essential genes in eukaryotes using protein sequence-derived features. *Comput Struct Biotechnol J* Elsevier 2019;**17**: 785–96.
85. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res* Oxford University Press 2006;**34**:564–74.
86. Chou K. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct Funct Bioinforma* Wiley Online Library 2001;**43**:246–55.
87. Jordan IK, Rogozin IB, Wolf YI, *et al*. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* Cold Spring Harbor Lab 2002;**12**: 962–8.
88. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* Elsevier; 2001;**305**:567–80.
89. Chen W-H, Trachana K, Lercher MJ, *et al*. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol Biol Evol* Oxford University Press 2012;**29**: 1703–6.
90. Wolf YI, Novichkov PS, Karev GP, *et al*. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci* National Acad Sciences 2009;**106**:7273–80.
91. Xu P, Ge X, Chen L, *et al*. Genome-wide essential gene identification in streptococcus sanguinis. *Sci Rep* Nature Publishing Group 2011;**1**:125.
92. Doyle MA, Gasser RB, Woodcroft BJ, *et al*. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. BMC genomics. *BioMed Central* 2010;**11**:222.
93. Goodacre NF, Gerloff DL, Uetz P. Protein domains of unknown function are essential in bacteria. *MBio Am Soc Microbiol* 2014;**5**:e00744–13.
94. Lu Y, Lu Y, Deng J, *et al*. Discovering essential domains in essential genes. *Methods Mol Biol* Springer 2015;**1279**: 235–45.
95. Yang J, Chen L, Kong X, *et al*. Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway. *PLoS One* Public Library of Science 2014;**9**:e107202.
96. Chen L, Zhang Y-H, Wang S, *et al*. Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways. *PLoS One* Public Library of Science 2017;**12**:e0184129.
97. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics [Internet]* BioMed Central 2009 [cited 2018 May 17];**10**:290. Available from. http://www.biomedcentral.com/1471-2105/10/290.
98. Peng C, Gao F. Protein localization analysis of essential genes in prokaryotes. *Sci Rep* Nature Publishing Group 2014;**4**:6001.
99. Akerley BJ, Rubin EJ, Novick VL, *et al*. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci. *National Acad Sciences* 2002;**99**:966–71.
100. Jeong H, Oltvai ZN, Barabási A-L. Prediction of protein essentiality based on genomic data. *ComPlexUs* Karger Publishers 2002;**1**:19–28.
101. Jacobs MA, Alwood A, Thaipisuttikul I, *et al*. Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proc Natl Acad Sci* National Acad Sciences 2003;**100**: 14339–44.
102. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* Cold Spring Harbor Lab 2002;**12**:37–46.
103. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* Oxford University Press 2005;**22**:803–6.
104. Joy MP, Brock A, Ingber DE, *et al*. High-betweenness proteins in the yeast protein interaction network. *Biomed Res Int* Hindawi 2005;**2005**:96–103.
105. Wang J, Peng W, Wu F-X. Computational approaches to predicting essential proteins: a survey. PROTEOMICS. *Clin Appl [Internet]* 2013 [cited 2018 May 17];**7**:181–92. Available from. http://www.ncbi.nlm.nih.gov/pubmed/23165920.
106. Wuchty S, Stadler PF. Centers of complex networks. *J Theor Biol* Elsevier 2003;**223**:45–53.
107. Zhang X, Acencio ML, Lemke N. Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review. *Front Physiol* Frontiers 2016;**7**:75.
108. Bonacich P. Factoring and weighting approaches to status scores and clique identification. *J Math Sociol* Taylor & Francis 1972;**2**:113–20.
109. Mariani MS, Ren Z-M, Bascompte J, *et al*. Nestedness in complex networks: observation, emergence, and implications. *Phys Rep* Elsevier 2019;**813**:1–90.
110. Koschützki D, Schreiber F. Comparison of centralities for biological networks. *Ger Conf Bioinforma*. *Citeseer*; Berlin 2004. p. 199–206.
111. Hwang Y-C, Lin C-C, Chang J-Y, *et al*. Predicting essential genes based on network and sequence analysis. *Mol Biosyst* Royal Society of Chemistry 2009;**5**:1672–8.

112. Yeh I, Hanekamp T, Tsoka S, *et al*. Computational analysis of plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res* Cold Spring Harbor Lab 2004;**14**:917–24.
113. Rahman SA, Schomburg D. Observing local and global properties of metabolic pathways:'load points' and 'choke points' in the metabolic networks. *Bioinformatics* Oxford Univ Press 2006;**22**:1767–74.
114. Xiao N, Cao D-S, Zhu M-F, *et al*. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* Oxford University Press 2015;**31**:1857–9.
115. Zhu M, Dong J, Cao D-S. rDNAse: R package for generating various numerical representation schemes of DNA sequences. 2016;
116. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, *et al*. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* Oxford University Press 2017;**33**:3387–95.
117. Muhammod R, Ahmed S, Md Farid D, *et al*. PyFeat: a python-based effective feature generation tool for DNA, RNA and protein sequences. *Bioinformatics* Oxford University Press 2019;**35**:3831–3.
118. Peden J. Analysis of codon usage (Doctoral dissertation, University of Nottingham, Nottingham, England). 2000. Retrieved from http://citeseerx.ist.psu.edu/.
119. Hagberg A, Swart P, S Chult D. *Exploring network structure, dynamics, and function using NetworkX*. *Los Alamos National Lab.(LANL)*. NM (United States: Los Alamos, 2008.
120. Yip KY, Yu H, Kim PM, *et al*. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* Oxford University Press 2006;**22**:2968–70.
121. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. *BioMed Central* 2008;**9**:559.
122. Savojardo C, Martelli PL, Fariselli P, *et al*. BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res* 2018;**46**:W459–66.
123. Wu S, Zhu Y. ProPAS: standalone software to analyze protein properties. *Bioinformation* Biomedical Informatics Publishing Group 2012;**8**:167.
124. Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Sanromán M. Filter methods for feature selection–a comparative study. *Int Conf Intell Data Eng Autom Learn*. Berlin:Springer; 2007. p. 178–87.
125. Hui KH, Ooi CS, Lim MH, *et al*. An improved wrapper-based feature selection method for machinery fault diagnosis. *PLoS One* Public Library of Science San Francisco, CA USA 2017;**e0189143**:12.
126. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics* Hindawi 2015;**2015**:198363.
127. Cheng J, Wu W, Zhang Y, *et al*. A new computational strategy for predicting essential genes. *BMC Genomics [Internet]* 2013 [cited 2018 May 17];**14**:910. Available from. http://www.ncbi.nlm.nih.gov/pubmed/24359534.
128. Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–57.
129. He H, Bai Y, Garcia E, *et al*. Adaptive synthetic sampling approach for imbalanced learning. *Proc Int Jt Conf Neural Networks* 2008. pp. 1322–1328.
130. Cheng J, Xu Z, Wu W, Zhao L, Li X, Liu Y, *et al*. Training Set Selection for the Prediction of Essential Genes. Kaderali L, editor. PLoS One [Internet]. 2014 [cited 2018 May 17];**9**:e86805. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24466248
131. Nigatu D, Sobetzko P, Yousef M, *et al*. Sequence-based information-theoretic features for gene essentiality prediction. *BMC Bioinformatics* Springer 2017;**18**:473.
132. Tian D, Wenlock S, Kabir M, *et al*. Identifying mouse developmental essential genes using machine learning. *Dis Model Mech* Company of Biologists Ltd 2018;**11**: dmm034546.
133. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* IEEE 2013;**35**:1798–828.
134. Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. arXiv Prepr arXiv14062661. 2014 Jun 10.
135. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv Prepr arXiv13126114. 2013 Dec 20.
136. Van Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. *Int Conf Mach Learn PMLR* 2016;**48**:1747–56.
137. Qin Z, Johnsen R, Yu S, *et al*. Genomic identification and functional characterization of essential genes in Caenorhabditis elegans. *G3 Genes, Genomes, Genet* Oxford University Press 2018;**8**:981–97.
138. Saha S, Heber S. In silico prediction of yeast deletion phenotypes. *Genet Mol Res* 2006;**5**:224–32.
139. da Silva JPM, Acencio ML, Mombach JCM, *et al*. In silico network topology-based prediction of gene essentiality. *Phys A Stat Mech its Appl [Internet]*. North-Holland 2008 [cited 2018 May 17];**387**:1049–55. Available from. https://www.sciencedirect.com/science/article/pii/S0378437107011417.
140. Azhagesan K, Ravindran B, Raman K. Network-based features enable prediction of essential genes across diverse organisms. Mande SC, editor. *PLoS One [Internet]* 2018 [cited 2019 Nov 30];**13**:e0208722. Available from. http://dx.plos.org/10.1371/journal.pone.0208722.
141. Zhang X, Xiao W, Hu X. Predicting essential proteins by integrating orthology, gene expressions, and PPI networks. *PLoS One Public Library of Science* 2018;**13**:e0195410.
142. Torgo L. *Data mining with R: learning with case studies*. London:CRC press, 2016.
143. Olson DL, Delen D. *Advanced data mining techniques*. Berlin:Springer Science & Business Media, 2008.
144. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta (BBA)-Protein Struct* Elsevier 1975;**405**:442–51.
145. Smeeton NC. Early history of the kappa statistic. *Biometrics* 1985;**41**:795.
146. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. Brock G, editor. *PLoS One [Internet]* 2015 [cited 2019 Nov 30];**10**:e0118432. Available from:. http://dx.plos.org/10.1371/journal.pone.0118432.
147. Manimaran P, Hegde SR, Mande SC. Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages. *Mol Biosyst Royal Society of Chemistry* 2009;**5**:1936–42.
148. Tong X, Campbell JW, Balázsi G, *et al*. Genome-scale identification of conditionally essential genes in E. coli by

DNA microarrays. *Biochem Biophys Res Commun* Elsevier 2004;**322**:347–54.
149. D'Elia MA, Pereira MP, Brown ED. Are essential genes really essential? *Trends Microbiol* Elsevier 2009;**17**:433–8.
150. Al-Khatib AM. Cooperative machine learning method. *World Comput Sci Inf Technol J(WCSIT)* 2011;**1**:380–3.
151. Goodfellow I. Nips 2016 tutorial: generative adversarial networks arXiv Preprint arXiv170100160. 2016 Dec 31.
152. Thurmond J, Goodman JL, Strelets VB, *et al*. FlyBase 2.0: the next generation. *Nucleic Acids Res* Oxford University Press 2018;**47**:D759–65.
153. Bult CJ, Blake JA, Smith CL, *et al*. Mouse genome database (MGD) 2019. *Nucleic Acids Res* Oxford University Press 2019;**47**:D801–6.
154. Cherry JM, Hong EL, Amundsen C, *et al*. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res* Oxford University Press 2012;**40**: D700–5.
155. Wagner J, Baur T, Zhang Y, *et al*. Applying cooperative machine learning to speed up the annotation of social signals in large multi-modalcorpora. arXiv. preprint arXiv:1802.02565. 2018 Feb 7.