Jaworsky *et al.*
Health Information Science and Systems (2023) 11:54
https://doi.org/10.1007/s13755-023-00254-7

Health Information Science and Systems

METHODOLOGY

Interrelated feature selection from health surveys using domain knowledge graph

Markian Jaworsky[1*], Xiaohui Tao[1], Lei Pan[2], Shiva Raj Pokhrel[2], Jianming Yong[1] and Ji Zhang[1]

Abstract

Finding patterns among risk factors and chronic illness can suggest similar causes, provide guidance to improve healthy lifestyles, and give clues for possible treatments for outliers. Prior studies have typically isolated data challenges from single-disease datasets. However, the predictive power of multiple diseases is more helpful in establishing a healthy lifestyle than investigating one disease. Most studies typically focus on single-disease datasets; however, to ensure that health advice is generalized and contemporary, the features that predict the likelihood of many diseases can improve health advice effectiveness when considering the patient's point of view. We construct and present a novel knowledge-based qualitative method to remove redundant features from a dataset and redefine the outliers. The results of our trials upon five annual chronic disease health surveys demonstrate that our Knowledge Graph-based feature selection, when applied to many machine learning and deep learning multi-label classifiers, can improve classification performance. Our methodology is compatible with future directions, such as graph neural networks. It provides clinicians with an efficient process to select the most relevant health survey questions and responses regarding single or many human organ systems.

Keywords: Feature selection, Risk factors, Knowledge graphs, Chronic illness

Introduction

Various risk factors can be used as predictor variables in the likelihood of developing chronic diseases. With awareness, patients can adapt their lifestyles to improve their chances of long-term survival. Risk factors can be categorized as lifestyle, environmental, or biomedical and can change over time. The WHO classifies diseases, injuries, and causes of death into 17,000 unique codes. Thousands of predictive models for each of the individual scenarios with a ranking of features is more than any patient requires to practice a healthy lifestyle. Therefore, a single model that best predicts many chronic diseases is the best-summarized information that can be given to the general public to achieve realistic lifestyle change.

Six known cancer types are positively correlated with a diagnosis of diabetes. According to the 2020 study by Wang et al. [1], diabetes increases the risks of multiple cancer diagnoses and outcomes, including pancreatic, liver, colorectal, breast, endometrial, and bladder. The characteristics of diabetes, high sugar, and insulin levels, with inflammation, are also known risk factors for cancer cells to proliferate, grow, and metastasize. This understanding helps patients determine their cancer risk as a combination of exposure to multiple risk factors. Risk factors trigger DNA damage and cause inflammation in the human body, promoting the lifespan of cancer cells.

Unfortunately, few frameworks are designed to address the multiple combined challenges, despite many research papers that aim to solve data challenge scenarios when analyzing health survey data. Jing et al. [2] proposed a novel feature selection approach to handling high-dimensional imbalanced class data when designing a classifier to address the issue of data nonlinearity. The shortcomings of this particular study are the lack of preparation of a data set containing missing values and the handling of outliers. The study of Zhang et al. [3] highlights that electronic medical records (EMR) have many inconsistent formats, and the range formatting continues to increase as new devices are

*Correspondence: Markian.Jaworsky@usq.edu.au
[1] School of Mathematics, Physics, and Computing, University of Southern Queensland, Toowoomba, QLD, Australia
Full list of author information is available at the end of the article

invented to produce EMR data. Furthermore, the contribution made by Vimalachandran et al. [4] exposes new privacy and security risks when novel devices are introduced to create new EMR data. Huang & Liu [5] provided the insight that traditional feature selection algorithms assume that features are unrelated and only consider the strength of the relationship between a predictor variable and the response variable. There is an opportunity to determine whether the algorithms for selecting features can be improved by considering which features have the strongest interrelations.

Large language models (LLMs) such as GPT and BERT have demonstrated research success and progress in handling unstructured text for various tasks across many industries. The Agrawal [6] thesis on the creation of structured information from clinical text using LLM is challenged as raw medical notes lack entity information and creative solutions are required to derive entity-to-entity relations between illnesses, patients, and treatments.

In this paper, we attempt to answer the research question: *How can we introduce innovation in identifying data patterns in classifying cancer and its subtypes?* We propose using a knowledge graph to exploit the structure of the chapters of the International Classification of Diseases (ICD) code published by the World Health Organization (WHO). By constructing a knowledge graph of features that pass a Wilcox Rank significance test, we aggregate a score to identify the most interrelated features of our datasets and filter our features for selection based on their rank. Our approach to knowledge graph construction automation differs from most research using statistics of words and their frequency of usage in chapters that describe human body organ systems.

Our approach selects many significant interrelated features from health survey data based on their question text frequencies in the WHO ICD codes. Based on multi-label classification trials in health surveys published by the Behavioral Risk Factor Surveillance System (BRFSS), we recommend selecting the knowledge graph's top 5% to 75% features according to their aggregate score. Figure 1 illustrates the optimal percentage of features to be selected from each BRFSS annual survey from 2017 to 2021. The Y-Axis provides a range from 0% to 75%. The blue bars indicate the annual optimum percentage of features identified for classification performance, with the points on the blue line indicating the actual value for that given year. Our approach gives insight into features with predictive power while reducing the number of features subject to missing values and effectively increasing the final dataset sample size.

The contributions of this paper are:

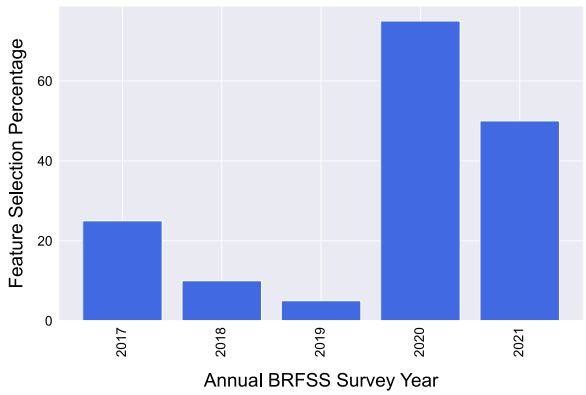

Fig. 1 KG-based feature selection—optimum percentage of features selected. The bar chart indicates the annual optimum percentage of features identified for classification performance

- performing feature selection based on an aggregated scoring method of the most interrelated features in a knowledge graph,
- demonstrating the performance improvement of multi-label chronic illness by selecting the most interrelated features in a dataset,
- provides future directions of knowledge-based feature relations applying our approach to determine feature relations.

The remaining sections of this paper are as follows: Presentation of background on feature selection and knowledge graphs in Sect. 2; Presentation of the framework for our research design and relevant datasets; Presentation of our evaluation of the multi-label algorithms in Sect. 4; Discussion of our research results in Sect. 5; Proposal hypothesis that relations described as linear ordinal variables may have more predictive power than nominal variables in our future directions in Sect. 6, and finally, our conclusion in Sect. 7.

Related work

Gaps in existing research

Parekh & Fahim [8] used the BRFSS health survey to study and construct ML predictive models for marijuana usage. They claimed that the task of selecting features from a health survey dataset is constrained by standard statistical methodology and that novel approaches are required. Chen & Wu [9] explain their main research objective in their study of Lung Cancer risk factors: identifying new causative factors of the disease to improve early detection.

In addition to eliminating redundant features, techniques can remove features for which any relation with a response variable has no explanation, preventing

spurious correlations via knowledge graph-driven feature selection. An overview of the model overfitting and proposed solution by Ying et al. [11] explains that the inclusion of irrelevant features in the training of a model is the cause of overfitting and is realized when a trained model performs the task of classification poorly on previously unseen data. To Handle overfitting, we can perform feature selection before model training or introduce a dropout layer, which divides model features into subsets and incrementally drops irrelevant features.

The study of Kim [12] explains that feature selection typically overlooks how features are interrelated, susceptible to data loss, and does not use labels to reduce datasets to a lower-dimensional representation of the original format. The Jaworsky et al. [13] knowledge graph proposes the ability to determine how features in a dataset may be interrelated. The proposal exploits word frequencies as they appear in the structure of chapters of the knowledge contained in the WHO ICD. Using a Wilcox Rank significance test for each pairing of features, only the responses to the significantly correlated health survey questions are used for the knowledge graph output.

For the classification of type 2 diabetes, Howlader et al. [18] employed several feature selection methods that have previously been proven effective when used in conjunction with machine learning. The result of the feature selection methods was the listing of candidate feature subsets to trial with machine learning classifiers. Ultimately the features that gave the best prediction outcomes were identified and ranked. This method of identifying features is resource-intensive, time-consuming, and yet not guaranteed to find an optimum feature subset. For data mining, relevant cause of death (COD) features from lung cancer Deng et al. [7] propose a random forest (RF) model of 10 selected features. RF and multinomial logistic regression (MLR) were chosen as candidates for classification, partly due to a prior similar study of breast cancer which determined that RF outperformed support vector machine (SVM) and artificial neural network (ANN) in the task of multi-category COD. The selection of features was ranked and incrementally tested as a group of numbers of features to determine an optimal model, which found that the model peaked with 10 features. The accuracy was reduced after 10 features were included. It is essential to realize that an optimum number of features to use for classification is specific to the dataset from which the features are derived. Our study gives recommendations regarding the percentage of features as opposed to an aggregated number. For the classification of skin lesions, Akram et al. [19] proposed a novel technique of fusing features to reduce the dimensions, retain their overall predictive power, and perform feature extraction. The proposed method achieves improved levels of classification performances against baseline measures and ultimately reduces the footprint of the selected feature subset to a minor portion of the original size. The study suggests testing the methodology on more datasets. Still, it does not address the question of explainable selection of features that specifically hold the predictive power of the response variables.

These studies have achieved success in predicting a single disease outcome. Multi-label classification is required to map features to many chronic illnesses or illness subtypes. Binary relevance multi-label classifiers are most common for this scenario. However, limitations exist in deriving semantic and contextual feature relations as noted by Nam et al. [20]. Binary relevance-based models may achieve optimum levels of accuracy for some individual labels; however, this does not equate to an optimum level of accuracy for all labels. Waegeman et al. [21] explains that binary relevance classification of individual labels leads to a broader range of diverse features. However, joint label feature selection can result in increased prediction accuracy. The additional benefit of a stacked binary relevance is an explainable set of predictor variables and a single classifier that is less prone to overfitting. For simplification of multi-label evaluation, it is possible to compress the combinations of labels into a single binary value by applying a maximum function to each classifier prediction. By using a Binary relevance label, which is the simplest of approaches for multi-label classification, as compared by Madjarovet al. [22], a (macro) average precision, recall, and F1-score can be derived for assessing the classifier performance.

Current state-of-the-art

An important consideration in classification performance evaluation is to ensure that both majority and minority classes have equal levels of precision, ensuring that training and test samples are adequately selected. Without an adequate approach to handling imbalanced data, classifiers will be biased toward majority classes to obtain favorable accuracy results [26–28]. Cross-validation is a common technique in data science, where training and test data are resampled into multiple batches, helping prevent overfitting of the classifier [29]. In the Gonzalez-Dias et al. [30] review of methods for determining the effectiveness of vaccine immune responses and potential side effects, the importance of measuring both the sensitivity and specificity of predicting the vaccine response is highlighted. The study suggests that the harmonic mean of the F1-score and confusion matrix should be used instead of the accuracy of outcome values being unbalanced. The most common surveyed multi-label classification evaluation

method is the F1-score or alternatively named harmonic mean performance metric [5, 10, 12, 26, 27, 30–40].

The study of extreme classification involves robust statistical modeling in many classes. However, in most cases, a classifier will rarely predict all labels seen during training, defined as a long-tail distribution. Bengio et al. [41] highlight that famous classifier performance metrics Hamming loss and 0–1 loss are unsuitable for extreme classification. Instead, the F-measure and precision of k labels ($P@k$) metrics are the popular choices, and there is still an opportunity to implement an improved measure that gives the most coverage of sparse labels.

Knowledge graphs are proposed as an alternative to resampling and upsampling techniques in classifying rare diseases from extremely imbalanced class datasets. The benefit of resampling methods is that they are packaged in software libraries and are very quick and easy to implement. However, the same can not be said of knowledge graphs, which are time-consuming and resource-intensive. Li et al. [14] explains that resampling, data synthesis, and cost-sensitive learning are typical imbalanced class handling approaches. However, rare disease prevalence can exceed 1:1000, where typically class imbalance range is 1:4. Knowledge graphs can benefit the classification of rare diseases even if not all knowledge is captured. Tao et al. [15] proposed creating a knowledge graph representation of the National Health and Nutrition Examination Survey (NHANES) to better data mining potential for chronic illness relationships. Medical domain knowledge graph and knowledge-based methods improved classification results of multi-label patient health status studies in Pham et al. [16, 17] compared to baseline approaches.

Several recent contributions have been made following the Open Graph Benchmark (OGB) for graph-based neural network node and edge property predictions as summarised in the paper of Hu et al. [23]. In addition to standardizing a format for structuring graph nodes, features, edges, and other graph metadata, the OGB framework has an online platform where datasets are published for further research. Apart from OGB, which is based upon the PyTorch Geometric library, Reiser et al. [24] claim that other noteworthy graph neural network libraries are Deep Graph Library (DGL), which also uses the PyTorch libraries, and Spektral/StellarGraph which use the TensorFlow-Keras framework. Despite the promising capability of graph neural networks (GNN), Li et al. [25] claim that experts find usage of GNN challenging due to the requirement of the combined programming, machine learning, and graph modeling skills. Another challenge for graph neural networks is computing power and reliance on a graphics processing unit (GPU), which is not typically available on a standard computer.

Novelty of our research

Feature selection and extraction methods are typically used to reduce dimensionality, complexity, and redundancy to ensure that the feature engineering process of building predictive models is an efficient, explainable process that produces optimum and repeatable classification results. However finding the optimal subset of features to perform chronic illness multilabel classification is an exhausting computing exercise that requires significant computing resources, involves pairing every feature subset with a target label, and individually assesses every combination of features until optimal classification performance is achieved. Alternatively, industry expertise is called upon to select features for predictive modeling, but little is gained in finding new causes and correlating attributes of chronic illness, in order to improve the time taken to diagnose patients and improve patient survival outcomes. Neural networks automate the process of feature subsetting and modeling but give little insight as to which feature permutations have the best classification performance, and are also subject to requiring significant computing resources, and design configurations that are least susceptible to overfitting.

Our unique knowledge graph-based approach to feature selection bypasses the individual pairing of features and feature subsets with a target label and its assessment of predictive capability. Instead, the most related features in the dataset are selected with a user-defined percentage of features to be selected. This process saves computing time, requires less computing power, and provides a transparent trail of features used in a predictive model. This novel approach is also adaptable to graph neural network modeling where relations between nodes and node features are relevant in the predictive capability of the architecture. Using low computing resources in this paper, we provide explainable results and ensure that the features selected are significantly correlated and relevant to the field of study.

Problem statement and data

- Diseases have their data mining patterns, but to predict and classify multiple diseases, we must identify the many data patterns and overcome their challenges.
- Better understanding datasets with many diseases will ensure health advice is relevant, and we must isolate those features and their relations that can predict the likelihood of chronic illness.
- Identifying predictor features with a 1-to-many cancer relationship improves health advice better than predictor features with only 1-to-1 relationships.

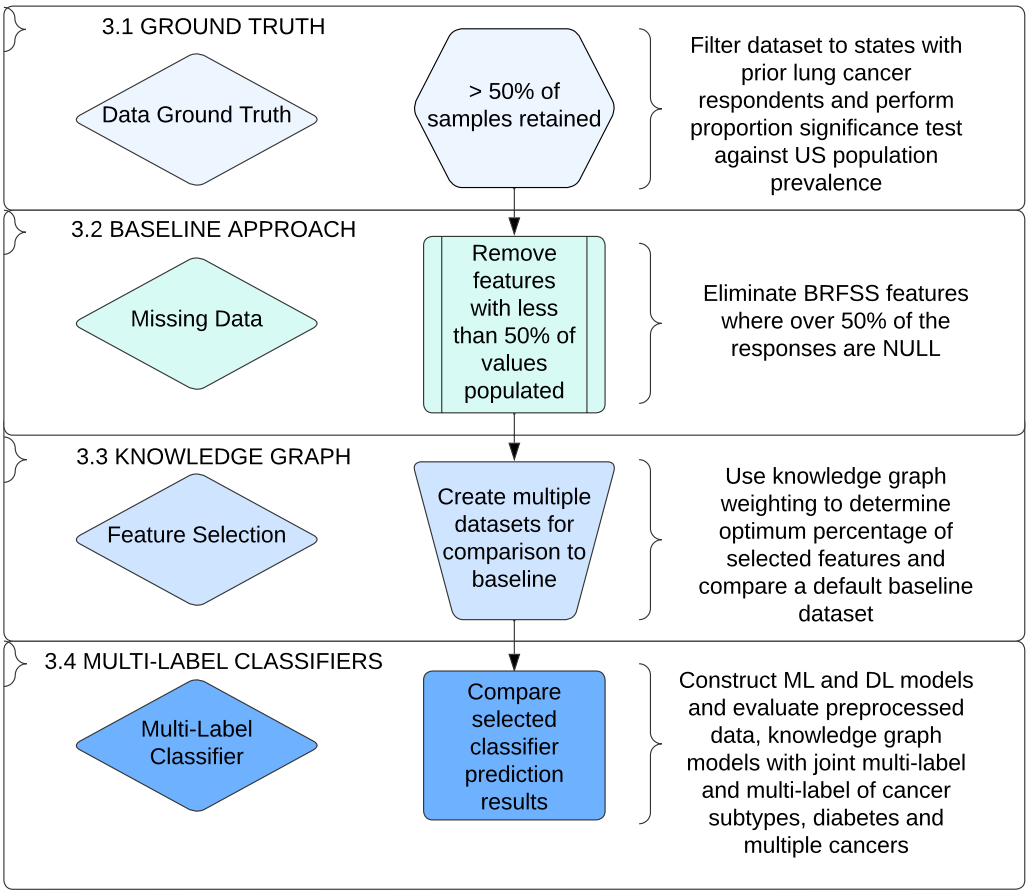

Fig. 2 Research design—details of our proposed methodology. Our step-by-step process for data preparation, feature selection and handling of missing values, and the selection of multi-label classifiers

The United States CDC anonymized annual BRFSS survey data across all 50 U.S. states from 1985 to 2021.[1] Respondents include those with and without diagnosed chronic health conditions. The questionnaires are designed to establish behaviors and risk factors associated with chronic health conditions to prevent the incidence of those conditions proactively. A typical annual BRFSS health survey consists of survey questions and answers, where each answer is represented by a nominal variable, which is a typical 'yes', 'no', 'do not know', 'refused to answer', or a numerical value to indicate multiple instances.

The 11th WHO revision of ICD codes[2] consists of a total of 1.6 million categorical and conceptual clinical terms. An online tool enables a search function to determine different sub-types of lung cancer, associated symptoms and risk factors, and likely stages of cancer progression. By extracting health survey question words and filtering common stop words, each question word can be cross-referenced with the 26 WHO ICD Code chapters specific to the knowledge of all documented human diseases. From the question text chapter frequencies, we can obtain a novel method for determining the strength of relations between health survey question features. The benefits of using the WHO ICD codes to construct a knowledge graph are the completeness of the data set that has slowly evolved over the past century and the logical placement of diseases.

Proposed methodology

We employ a modular approach in our research and subdivide the research problem into several components to resolve them independently using different methods. A high-level view of our modular approach is shown in Fig. 2, which demonstrates the step-by-step process for data preparation, feature selection and handling of missing values, and the selection of multi-label classifiers. We explain each of them in detail in the following.

[1] https://www.cdc.gov/brfss/annual_data/annual_2021.html

[2] https://www.who.int/standards/classifications/classification-of-diseases

Table 1 Ground truth—insignificant proportion comparison p-value = 0.05129

Dataset	Lung Cancer prevalence	Candidate population	Proportion
BRFSS 2020 State-Based Subset	440	224,650	0.001958602
USA General Population 2018	582631	326800000	0.001782837

Ground truth analysis

The baseline dataset to be used first applies a filter of the geographical state of the respondent, and states eliminated from the dataset have been identified by searching through prior years of surveys and deselecting states without lung cancer candidates. To establish a dataset representative of the ground truth, the BRFSS annual survey can be filtered by states until an insignificant proportion of lung cancer, a chronic disease with a high prevalence amongst both male and female patients and comparable to the known prevalence in the general population.

The US Cancer Institute estimated that the prevalence of lung cancer in 2018 was 582,631 living people[3]. In the same year, the U.S. Census Bureau[4] estimated the US national population to be 326.8 million. By filtering the BRFSS 2020 Annual Survey to the subset of states with lung cancer patients recorded in the 11 years of 2010 to 2020, it is possible to achieve an insignificant proportion of lung cancer prevalence in our source data set for analysis, compared to the general population of the United States. The results of the geographically based method of filtering data by states are displayed in Table 1. Secondly, we remove the features where over 50% of the values are missing (blank).

Segregation of baselines

Pham et al. [16] proposed a novel knowledge graph to classify multiple chronic diseases and applies a standard method for preparing health survey data, only using characteristics where more than 50% of the values exist. If the required values of selected features are missing, samples are removed from the dataset. We regard this as our baseline dataset. However, for our feature-selected datasets, we apply a selection of features where the missing values are less than the frequency 50% and only select features where they appear in the top quantile of our knowledge graph, which is determined by the aggregated edge score, multiplied the number of related features. We then compare the performance of multi-label classifiers between the baseline and the feature-selected datasets to determine if the feature selection algorithm provides improvement. Dinh et al. [42] predicted both diabetes and cardiovascular disease from the NHANES survey using responses that were not less than 50%

Many machine learning-based studies use popular large-scale health surveys, such as the BRFSS Annual Health Survey, to discover predictor variables and help predict chronic disease [9, 43, 44]. Our framework's novelty and demonstrated benefit improve the performance of classifiers using knowledge graph-based feature selection by identifying the most interrelated features. For our study, we re-use the baseline dataset preparation as trailed in our paper for nonlinear dataset transformation [45].

Figure 2 illustrates our step-by-step process for data preparation, feature selection and handling of missing values, and the selection of multi-label classifiers. Our experiments on related chronic illnesses for multi-label classification are based on the most significant annual health survey but can be applied to any health survey, and our knowledge graph is constructed from the world's oldest and most comprehensive knowledge base of human disease.

The labels of our selected datasets for diabetes and cancer are converted into a binary format so that the values of 1 or 2 denote malignant diagnosis, all other options are 0, and missing (blank) are omitted. For the cancer label, we convert to 0 if it is missing (blank), and any other recorded value to 1. With our multi-label configured to a binary representation of 0 or 1, evaluating classifier performance is simplified such that any predictions of a true positive of either diabetes or cancer can be interpreted as success. We are evaluating the difference between multi-label classifiers over baseline and feature-selected datasets.

Develop knowledge graph driven algorithm

Knowledge graphs have shown potential in personalized drug treatment, reported in the review of Zeng et al. [46]. Knowledge Graph-based algorithms help identify unstructured semantic relations between entities. The study notes that knowledge graphs should be measured for quality, and this quality metric should be used to guide maintenance and enhancements. A knowledge graph has a basic core structure as described by the Nicholson et al. [47] review of knowledge graphs for biomedical applications. To be meaningful, nodes and edges

[3] https://seer.cancer.gov/statfacts/html/lungb.html

[4] https://www.census.gov

are required to aid machine learning classification. Various approaches exist to construct knowledge that can be manually curated from the text in a time-consuming process, or automated processes can extract data from databases. Knowledge graph scalability and memory constraints can be encountered when training a classifier, and ensuring the completeness of the knowledge of a knowledge graph is the most common challenge.

Algorithm 1 lists the step-by-step requirements of transforming the health survey question list, which has linear and nonlinear responses, to a data set with significantly interrelated features selected, given the use of all the ICD chapters, or limited to specific chapters on human organs, or the domain knowledge is concentrated with that level of focus. Step 2 of the knowledge graph construction algorithm allows for the sequential decrement of percentage thresholds by single percentage points; in our study, we have opted to use percentage thresholds of the statistical quartiles 25%, 50%, 75%, 100%, as well as the standard p-value percentages of 5% and 10%. We aggregate the number of feature interrelations at Step 9 of the algorithm to determine an overall node ranking of the highest scores at Step 14, which specifies which features are selected when the percentage threshold at Step 2 is defined.

Algorithm 1 Threshold Setting & Feature Selection

Input: Raw BRFSS Health Survey Questions
Output: Selection of BRFSS Health Survey Questions

1: **procedure** KNOWLEDGE GRAPH FEATURE SELECTION
2: *Threshold* ← *100%*
3: *FeatureCount* ← *Number of DataSet Features*
4: **while** *Threshold* ≥ *1%* **do**
5: for Each Knowledge Graph Row:
6: for Each Knowledge Graph Feature:
7: *Weight* ← *Edge Value*
8: **if** *Row Feature Repeat* **then**
9: *Weight* ← *Weight* + *Edge Value*
10: **end if**
11: **end for do**
12: **end for do**
13: **end while**
14: *FeatureList* ← *Sort Features By Highest Aggregated Weight*
15: *SelectionNumber* ← *FeatureCount* * *Threshold Percentage*
16: *FeatureSelection* ← *Top SelectionNumber in FeatureList*
17: *Threshold* ← *ThresholdNumber - 1%*
18: **end procedure**

Table 2 illustrates the knowledge graph nodes in the column Feature A, interrelated features are listed in the column Feature B, and a significance test of the feature interrelations is displayed in the column of p values. At the same time, the column of Word count is the number of text words, two features measured for an interrelation, and also represents our edge scores of the knowledge graph. Our knowledge graph is implemented and available on GitHub[5].

Table 2 BRFSS 2021 knowledge graph—sample feature interrelations

Feature A	Feature B	p-value	Word count
1	38	0.049	17
1	50	0.047	14
1	80	0.038	15
1	126	0.038	15
1	129	0.047	17
1	131	0.035	15
1	136	0.042	18
1	139	0.042	14
1	140	0.049	13
1	141	0.049	13
1	145	0.044	16
1	165	0.036	17
1	169	0.039	19
1	185	0.047	17
1	193	0.045	18
1	206	0.030	15
1	213	0.042	15
..	..	Note: 5450 rows omitted due to size	..
228	209	0.042	9

Our feature selection method only looks at the input variables and does not evaluate the permutations and combinations of the feature subset against the target label. This unsupervised approach is novel and a simple and fast process for selecting features without requiring exhaustive computing resources. The task of selecting the most relevant health survey questions for predicting chronic illness can be performed in a few hours to aid medical professionals, leverage a knowledge graph of human organ systems to automatically determine relevant health survey questions specific to single or many human organ systems, depending on the type of chronic illness. Upon completion of the algorithm, the original health survey is reduced to a percentage portion of the original health survey, with only questions that have the significantly correlated keywords of other health survey questions remaining. Responses to these significantly correlated questions are then used to classify chronic diseases.

[5] https://github.com/mjaworsky/KnowledgeGraph

Table 3 Baseline vs feature selected BRFSS fivefold CV, macro precision, macro recall, macro F1-score, hamming loss—2017–2021 averages

Classifier	Baseline				Interrelated features			
	Precision	Recall	F1	HL	Precision	Recall	F1	HL
AB	0.532	0.500	0.478	0.097	0.670	0.512	0.502	0.074
CNN	0.276	0.500	0.304	0.086	0.383	0.500	0.412	**0.064**
KNN	0.474	0.500	0.474	0.096	0.464	0.500	0.480	0.076
LDA	0.508	0.500	0.478	0.096	**0.814**	0.610	0.592	0.074
LR	0.480	0.500	0.474	0.096	0.574	0.504	0.484	0.077
MNB	0.486	0.510	0.442	0.307	0.490	0.516	0.480	0.141
RB	0.516	0.506	0.488	0.114	0.748	**0.672**	**0.670**	0.067
RF	0.452	0.500	0.474	0.149	0.566	0.504	0.486	0.076
SVM	0.452	0.500	0.474	0.096	0.464	0.500	0.480	0.077

Bold values indicate the optimum result by column

Multi-label classifiers

The prior success with other health survey studies has guided our selection of classifiers. Machine learning prediction of mortality for children under 5 years of age, from 2016 by Bitew et al. [48], used logistic regression (LR), RF, and K-nearest neighbors (KNN) classifiers. Using multiple metrics, the study suggests that RF had the greater predictive power of the three algorithms. To predict Parkinson's staging, Prashanth and Roy [49] trial the imbalanced data classifiers RUSBoost and Adaboost, along with SVM, RF, LR, closest neighbors K, neutral approaches, negative approaches, and deep learning classification models. The study noted that the data contained imbalanced classes and nonlinearity. In the scenario, the deep learning classification obtained the most powerful classification results on average.

Ricciardi et al. [50] applied a linear discriminant analysis (LDA) classifier in conjunction with the principal component analysis (PCA) feature extraction algorithm to predict illness in a data set constructed by clinicians. The study results demonstrated the value of eliminating redundant features before classification. For the classification of specific text sentiment, Georgakopoulos et al. [51], constructed a Convolutional Neural Network (CNN) classifier. The study explains that the CNN model consists of multiple layers and relies on backpropagation to learn and improve predictive power. The study compares the CNN model to other machine learning algorithms and finds that a 3-layer CNN model, with a learning rate of 0.005, outperforms the machine learning models in the task of toxic text classification.

In stark contrast to existing work [48–51], we develop and analyze a binary relevance classifier for a response variable with a combination of diabetes, cancer, or both, which can exploit the correlation between the two chronic diseases. We believe that identifying predictor variables with a 1-to-many cancer relationship can improve health advice by only identifying individual predictor variables with 1-to-1 relationships to chronic illnesses. As individual diseases have unique data patterns, our classifier always seeks to predict multiple diseases, as we aim to overcome the data challenges that occur when attempting to predict the outcome.

Performance evaluation, findings, and results

Multi-label classifier performance evaluation

Since there is a verified link between these two chronic diseases, we have created a binary relevance classifier to explore and exploit the correlation for a response variable that includes diabetes, cancer, or both. Instead of focusing on finding individual predictor variables with one-to-one correlations with chronic diseases, our team sees that it would be beneficial to uncover predictor variables with a one-to-many cancer association. Our algorithm can predict several diseases since we have attempted to address the data problems that arise when predicting each specific condition.

The summary of results listed in Table 3 compares the baseline dataset against the optimum feature selected dataset, using the measures of precision, recall, F1-score, and Hamming loss. The classifier acronyms in Tables 3 are interpreted as follows:

- AB = AdaBoost
- CNN = Convolutional Neural Network
- KNN = K-Nearest Neighbours
- LDA = Linear Discriminant Analysis
- LR = Logistic Regression
- MNB = Multinomial Naive Bayes
- RB = RUSBoost
- RF = Random Forest
- SVM = Support Vector Machine

Table 4 Significance test BRFSS fivefold CV, macro precision, macro recall, macro F1-score, Hamming loss—2017–2021 averages

Metric	Baseline	Our method	Mean diff	t	df	p-value
Macro Precision	0.464	0.575	−0.111	−3.140	8	**0.014**
Macro Recall	0.502	0.535	−0.034	−1.651	8	0.137
Macro F1-Score	0.454	0.510	−0.056	−2.621	8	**0.031**
Hamming Loss	0.126	0.081	0.046	2.815	8	**0.023**

Table 5 Baseline vs interrelated features BRFSS fivefold CV, macro precision, macro recall, macro F1-Score, hamming loss—top result

Metrics	%	AB	CNN	KNN	LDA	LR	MNB	RB	RF
Precision	5%	0.440	0.310	0.440	0.440	0.440	0.500	**0.930**	0.560
	10%	0.840	0.280	0.440	0.850	**0.870**	0.500	**0.870**	0.890
	25%	0.720	0.270	0.500	**0.950**	0.500	0.500	0.570	0.500
	50%	0.830	0.300	0.490	0.900	**0.930**	0.500	**0.930**	0.490
	75%	0.840	0.250	0.500	**0.920**	0.840	0.500	**0.920**	0.500
Recall	5%	0.500	0.500	0.500	0.500	0.500	0.500	**0.620**	0.500
	10%	0.550	0.500	0.500	0.520	0.510	0.500	**0.630**	0.520
	25%	0.500	**0.680**	0.500	0.500	0.500	0.500	0.610	0.500
	50%	0.510	0.500	0.500	0.520	0.500	0.500	**0.620**	0.500
	75%	0.500	0.500	0.500	**1.000**	0.510	0.530	**1.000**	0.500
F1	5%	0.470	0.330	0.470	0.470	0.470	0.370	**0.670**	0.460
	10%	0.560	0.290	0.470	0.510	0.470	0.410	**0.670**	0.490
	25%	0.480	0.290	0.550	0.470	0.500	0.500	**0.580**	0.500
	50%	0.480	0.330	0.500	0.500	0.460	0.500	**0.670**	0.500
	75%	0.460	0.280	0.500	**0.960**	0.480	0.470	**0.960**	0.500
Hamming	5%	0.111	**0.066**	0.111	0.111	0.111	0.118	0.145	0.111
	10%	0.112	**0.066**	0.112	0.112	0.112	0.126	0.111	0.112
	25%	**0.008**	0.066	**0.008**	0.009	**0.008**	0.107	0.038	**0.008**
	50%	0.010	0.066	**0.004**	0.011	0.010	0.110	0.013	0.010
	75%	0.009	0.028	0.009	**0.002**	0.010	0.143	**0.002**	0.009

Bold values indicate the optimum result by row

The comparisons in Table 3 show that for five consecutive BRFSS health surveys from 2017 to 2021, a data set with the WHO ICD-based knowledge graph feature selection of between 75% and 5% performs equally or more often better than the baseline data sets without the application of the feature selection method. Optimal classifier performance was obtained using the RUSBoost classifier in most years with a varying selection of percentages of feature candidates. In Table 4, we cascade all the values listed in Table 3, compare the difference between the average metric values for all of our ML and DL algorithm performances over the five years 2017–2021, and perform a significance test. These significance tests were performed using Student's paired t-test R programming function[6]. We observe an improvement in all 4 metrics used to measure classification performance, a significance test is useful against the measures of precision, recall, and F1-score as they skew towards understating the effectiveness of a new method [52].

Detailed results by year are accessible via Google Sheets online.[7]

[6] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

[7] https://docs.google.com/spreadsheets/d/e/2PACX-1vQOFp2s7H60fZNfUeF47NIOT4eN1UDEUq8wFJsPnV3MHZJd5i8LPVeR5FWwZ9SG3jmoRs837GXofliw/pubhtml

Ablation study

Table 5 provides 2017–2021, 5-year optimum sets of precision, recall, F1-scores, and Hamming loss of our study feature selection percentage thresholds. The results of this table show that the RUSBoost classifier is the most consistent performing across all 4 metrics. The performance of the LDA and LR classifiers indicates that the feature selection dimension reduction algorithm improves classification performance by eliminating nonlinear features.

Our implementation of Google TensorFlow CNN consists of 3 layers and dimensionality units per the count of features. The number of backpropagation epochs was increased from 10 to 100. Learning rates were decremented to 0.001 from 0.01 with the popular RELU activation function until the deep learning could predict a minority class label instance. This configuration level is a feasible uplift from default hyperparameter values, which does not require a graphics processing unit (GPU) hardware investment. Our general methodology for identifying the most significantly interrelated features from a health survey dataset before building a chronic illness classifier is a breakthrough approach as the operation can be reproduced with standard desktop computing resources. Additionally, it does not require exhaustive processing execution time to determine an optimal subset of features that may have been spuriously correlated or overfitted to a label in a disproportionate training set, but instead guided by actual domain knowledge.

Discussion

The general results of the RusBoost classifier for the 5 BRFSS surveys between 2017 and 2021 demonstrate that the approach to random sample reduction of the majority class improves the classifier in identifying and predicting minority class samples from highly imbalanced datasets [53]. Under-sampling of the benign (no diabetes and no cancer) major class has proven to be most effective on the dataset with feature selection applied. The low performance of the nonlinear and tree-based classifiers demonstrates that under-sampling has assisted the RUSBoost classifier in learning from the nonlinear variables. The linear classifiers have also performed reasonably well without under-sampling.

In fact, RUSBoost is a black-box ensemble model to handle both linear and nonlinear data, according to the RUSBoost classification performance explained by Carrasco et al. [54]. Furthermore, the RUSBoost implementation library[8] declares that the classifier is an ensemble of five algorithms (viz. Classification and Regression Tree, Decision Tree, RF, Naive Bayes, and SVM).

Conclusion

This study demonstrates that the construction of a knowledge graph can significantly improve feature selection in cancer-based health surveys by directly searching for interrelated characteristics. Constructing a health-based knowledge graph provides more transparency than automated deep-learning solutions by identifying significantly interrelated features. The selection of multi-label classifiers is essential, as health survey datasets consist of imbalanced classes and mixtures of linear and nonlinear variables. An ensemble of linear, nonlinear, and majority class undersampling multi-label classifiers can provide the best coverage for identifying and predicting true positive minority class samples. Our contribution of a knowledge graph-based feature selection method adds value to the RUSBoost and other classifiers by reducing features from high-dimensional datasets.

Our methodology proposes an efficient process to select significant interrelated health survey questions and responses with respect to the relevant human organ systems of chronic illness in the focus of their studies. Our methodology removes outliers and noisy features through a knowledge-based qualitative process, which caters to improving the context of feature selection in a health survey. A better understanding of feature interrelations not only improves feature selection but also provides the groundwork for discovering new links between features and expanding the relevancy of dataset features.

Future directions

The prevalence of linear variables in the BRFSS annual surveys is generally only a 20% ratio to nonlinear variables. A relevant future research question may ask whether linear variables should be weighted or preferred over nonlinear ones during feature selection. Trending GNN classification models also incorporate node and feature relations. The ability of a GNN architecture to combine raw datasets with node relations (edge metadata) can improve classification performance significantly. We hypothesize that node relations will provide more powerful information if they are represented as linear ordinal numerical variables instead of nonlinear nominal variables.

Furthermore, an improved understanding of risk factor feature relations can improve patient outcomes with early diagnosis, we can also discover new links between risk health survey factors. The ability to predict relations between keywords of the WHO ICD by human organ system can unlock clues for improving the health survey predictive power of chronic illness. In this paper,

[8] https://rdrr.io/cran/ebmc/man/rus.html

we explore the usage of feature relations described as a numeric significance p-value, between 2 significantly related features with an additional word count attribute. Due to the effectiveness of neural network architectures on high-volume data, there is an opportunity to describe the relations between 2 features in more detail and in terms of the number of keywords matched across each of the 26 chapters of the WHO ICD, expanding the dimensionality of the relationships and enhancing its predictive power.

Acknowledgements
Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2017-2021. International Classification of Diseases, Eleventh Revision (ICD-11), World Health Organization (WHO) 2019/2021 https://icd.who.int/browse11. Licensed under Creative Commons Attribution-No Derivatives 3.0 IGO license (CC BY-ND 3.0 IGO).

Declarations

Human Research Ethics
Human Research Ethics (HRE) application has been reviewed by the University (of Southern Queensland) Expedited Review process. The research proposal has been deemed to meet the requirements of the National Statement on Ethical Conduct in Human Research (2007), Human Research Ethics, University of Southern Queensland, Toowoomba, Queensland, 4350, Australia, Email: human.ethics@usq.edu.au

Author details
[1]School of Mathematics, Physics, and Computing, University of Southern Queensland, Toowoomba, QLD, Australia. [2]School of Information Technology, Deakin University, Waurn Ponds, VIC 3216, Australia.

Received: 21 January 2023 Accepted: 17 October 2023
Published: 16 November 2023

References

1. Wang M, Yang Y, Liao Z. Diabetes and cancer: epidemiological and biological links. World J Diabetes. 2020;11(6):227.
2. Jing X-Y, Zhang X, Zhu X, Wu F, You X, Gao Y, Shan S, Yang J-Y. Multiset feature learning for highly imbalanced data classification. IEEE Trans Pattern Anal Mach Intell. 2019;43(1):139–56.
3. Zhang Yong, Sheng Ming, Liu Xingyue, Wang Ruoyu, Lin Weihang, Ren Peng, Wang Xia, Zhao Enlai, Song Wenchao. A heterogeneous multimodal medical data fusion framework supporting hybrid data exploration. Health Inf Sci Syst. 2022;10(1):22.
4. Vimalachandran Pasupathy, Liu Hong, Lin Yongzheng, Ji Ke, Wang Hua, Zhang Yanchun. Improving accessibility of the Australian My Health Records while preserving privacy and security of the system. Health Inf Sci Syst. 2020;8:1–9.
5. Huang H, Liu H. Feature selection for hierarchical classification via joint semantic and structural information of labels. Knowl-Based Syst. 2020;195:105655.
6. Agrawal M. Towards scalable structured data from clinical text. PhD diss., Massachusetts Institute of Technology; 2023.
7. Deng F, Zhou H, Lin Y, Heim JA, Shen L, Li Y, Zhang L. Predict multicategory causes of death in lung cancer patients using clinicopathologic factors. Comput Biol Med. 2021;129:104161.
8. Parekh T, Fahim F. Building risk prediction models for daily use of marijuana using machine learning techniques. Drug Alcohol Depend. 2021;225:108789.
9. Chen Songjing, Sizhu Wu. Identifying lung cancer risk factors in the elderly using deep neural networks: quantitative analysis of web-based survey data. J Med Int Res. 2020;22(3):e17695.
10. Pan Liangrui, Ji Boya, Wang Hetian, Wang Lian, Liu Mingting, Chongcheawchamnan Mitchai, Peng Shaolaing. MFDNN: multi-channel feature deep neural network algorithm to identify COVID19 chest X-ray images. Health Inf Sci Syst. 2022;10(1):4.
11. Ying X. 2019, February. An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, No. 2, p. 022022). IOP Publishing.
12. Kim K. An improved semi-supervised dimensionality reduction using feature weighting: application to sentiment analysis. Expert Syst Appl. 2018;109:49–65.
13. Jaworsky M, Tao X, Yong J, Pan L, Zhang J, Pokhrel S. Automated knowledge graph construction for healthcare domain. In: Proceedings of the International Conference on Health Information Science, pp. 258–265, Springer; 2022.
14. Li X, Wang Y, Wang D, Yuan W, Peng D, Mei Q. Improving rare disease classification using imperfect knowledge graph. BMC Med Inform Decis Mak. 2019;19(5):1–10.
15. Tao X, Pham T, Zhang J, Yong J, Goh WP, Zhang W, Cai Y. Mining health knowledge graph for health risk prediction. World Wide Web. 2020;23:2341–62.
16. Pham Thuan, Tao Xiaohui, Zhang Ji, Yong Jianming, Li Yuefeng, Xie Haoran. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. Knowl-Based Syst. 2022;235:107662.
17. Pham Thuan, Tao Xiaohui, Zhang Ji, Yong Jianming. Constructing a knowledge-based heterogeneous information graph for medical health status classification. Health Inf Sci Syst. 2020;8:1–14.
18. Howlader KC, Satu MS, Awal MA, Islam MR, Islam SMS, Quinn J, Moni MA. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health Inf Sci Syst. 2022;10(1):2.
19. Akram T, Lodhi HM, Naqvi SR, Naeem S, Alhaisoni M, Ali M, Haider SA, Qadri NN. A multilevel features selection framework for skin lesion classification. Human-centric Comput Inf Sci. 2020;10:1–26.
20. Nam J. Learning Label Structures with Neural Networks for Multilabel Classification. PhD thesis, Technische Universität; 2019.
21. Waegeman W, Dembczyński K, Hüllermeier E. Multi-target prediction: a unifying view on problems and methods. Data Min Knowl Disc. 2019;33(2):293–324.
22. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. Pattern Recogn. 2012;45(9):3084–104.
23. Hu Weihua, Fey Matthias, Zitnik Marinka, Dong Yuxiao, Ren Hongyu, Liu Bowen, Catasta Michele, Leskovec Jure. Open graph benchmark: Datasets for machine learning on graphs. Adv Neural Inf Process Syst. 2020;33:22118–33.
24. Reiser Patrick, Eberhard André, Friederich Pascal. Graph neural networks in TensorFlow-Keras with RaggedTensor representation (kgcnn). Software Impacts. 2021;9:100095.
25. Li Mufei, Zhou Jinjing, Jiajing Hu, Fan Wenxuan, Zhang Yangkang, Yaxin Gu, Karypis George. Dgl-lifesci: an open-source toolkit for deep learning on graphs in life science. ACS Omega. 2021;6(41):27233–8.
26. Pes B. Learning from high-dimensional and class-imbalanced datasets using random forests. Information. 2021;12(8):286.
27. Liu M, Xu C, Luo Y, Xu C, Wen Y, Tao D. Cost-sensitive feature selection via f-measure optimization reduction. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017.
28. Xu D, Shi Y, Tsang IW, Ong Y-S, Gong C, Shen X. Survey on multi-output learning. IEEE Trans Neural Netw Learn Syst. 2019;31(7):2409–29.
29. Berrar D. Cross-validation; 2019.
30. Gonzalez-Dias P, Lee EK, Sorgi S, de Lima DS, Urbanski AH, Silveira EL, Nakaya HI. Methods for predicting vaccine immunogenicity and reactogenicity. Hum Vacc Immunother. 2020;16(2):269–76.
31. Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, Chen L, Qiu L. Exploratory study on classification of diabetes mellitus through a combined random forest classifier. BMC Med Inform Decis Mak. 2021;21(1):1–14.
32. Fan S-KS, Hsu C-Y, Jen C-H, Chen K-L, Juan L-T. Defective wafer detection using a denoising autoencoder for semiconductor manufacturing processes. Adv Eng Inform. 2020;46:101166.

33. Song D, Vold A, Madan K, Schilder F. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. Information Systems, p. 101718; 2021.
34. Gupta N, Bohra S, Prabhu Y, Purohit S, Varma M. Generalized zero-shot extreme multi-label learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 527–535; 2021.
35. Wang J, Zhou F, Wen S, Liu X, Lin Y. Deep metric learning with angular loss. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2593–2601; 2017.
36. Zhu L, Yang Y. Inflated episodic memory with region self-attention for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4344–4353; 2020.
37. Ibrahim MA, Khan MUG, Mehmood F, Asim MN, Mahmood W. Ghs-net a generic hybridized shallow neural network for multi-label biomedical text classification. J Biomed Inform. 2021;116:103699.
38. Melacci S, Ciravegna G, Sotgiu A, Demontis A, Biggio B, Gori M, Roli F. Domain knowledge alleviates adversarial attacks in multilabel classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence; 2021.
39. Zhu P, Wang H, Saligrama V. Learning classifiers for target domain with limited or no labels. In: Proceedings of the International Conference on Machine Learning, pp. 7643–7653, PMLR; 2019.
40. Ruas P, Neves A, Andrade VD, Couto FM, Aragon ME. Lasigebiotm at cantemist: Named entity recognition and normalization of tumour morphology entities and clinical coding of spanish healthrelated documents. In: IberLEF@ SEPLN, pp. 422–437; 2020.
41. Bengio S, Dembczynski K, Joachims T, Kloft M, Varma M. Extreme classification (dagstuhl seminar 18291). In: Dagstuhl Reports, vol. 8, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik; 2019.
42. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019;19(1):1–15.
43. Fang M, Chen Y, Xue R, Wang H, Chakraborty N, Su T, Dai Y. A hybrid machine learning approach for hypertension risk prediction. Neural Computing and Applications, pp. 1–11; 2021.
44. Zgodic A, Zahnd WE, Miller DP Jr, Studts JL, Eberth JM. Predictors of lung cancer screening utilization in a population-based survey. J Am Coll Radiol. 2020;17(12):1591–601.
45. Jaworsky M, Tao X, Yong J, Pan L, Zhang J, Pokhrel SR. Knowledge-Based Nonlinear to Linear Dataset Transformation for Chronic Illness Classification. Health Information Science. Lecture Notes in Computer Science, vol 14305. Springer, Singapore. HIS 2023.
46. Zeng X, Tu X, Liu Y, Fu X, Su Y. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol. 2022;72:114–26.
47. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020;18:1414–28.
48. Bitew FH, Nyarko SH, Potter L, Sparks CS. Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian demographic and health survey. Genus. 2020;76(1):1–16.
49. Prashanth R, Roy SD. Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning. Neurocomputing. 2018;305:78–103.
50. Ricciardi C, Valente AS, Edmund K, Cantoni V, Green R, Fiorillo A, Picone I, Santini S, Cesarelli M. Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Inform J. 2020;26(3):2181–92.
51. Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP. Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic conference on artificial intelligence, pp. 1–6; 2018.
52. Yeh, Alexander. More accurate tests for the statistical significance of result differences. arXiv preprint arXiv:cs/0008005 (2000).
53. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. Rusboost: Improving classification performance when training data is skewed. In: Proceedings of the 2008 19th International Conference on Pattern Recognition, pp. 1–4, IEEE; 2008.
54. Carrasco J, Lison F, Weintraub A. Rusboost: A suitable species distribution method for imbalanced records of presence and absence. A case study of twenty-five species of Iberian bats, bioRxiv; 2021.
55. Sahoo D, Liu C, Hoi SC. Malicious URL detection using machine learning: a survey. arXiv preprint arXiv:1701.07179; 2017.