applied sciences

MDPI

Article

High-Risk HPV Cervical Lesion Potential Correlations Mining over Large-Scale Knowledge Graphs

Tiehua Zhou , Pengcheng Xu, Ling Wang * and Yingxuan Tang

School of Computer Science, Northeast Electric Power University, Changchun Road, Jilin City 132011, China; thzhou@neepu.edu.cn (T.Z.); 2202101009@neepu.edu.cn (P.X.); 2202000690@neepu.edu.cn (Y.T.)
* Correspondence: smile2867ling@neepu.edu.cn

Abstract: Lesion prediction, a very important aspect of cancer disease prediction, is an important marker for patients before they become cancerous. Currently, traditional machine learning methods are gradually applied in disease prediction based on patient vital signs data. Accurate prediction requires a large amount and high quality of data, however, the difficulty in obtaining and incompleteness of electronic medical record (EMR) data leads to certain difficulties in disease prediction by traditional machine learning methods. Secondly, there are many factors that contribute to the development of cervical lesions, some risk factors are directly related to it while others are indirectly related to them. In addition, risk factors have an interactive effect on the development of cervical lesions; it does not occur in isolation, a large-scale knowledge graph is constructed base on the close relationships among risk factors in the literature, and new potential key risk factors are mined based on common risk factors through a subgraph mining method. Then lesion prediction algorithm is proposed to predict the likelihood of lesions in patients base on the set of key risk factors. Experimental results show that the circumvents the problems of large number of missing values in EMR data and discovered key risk factors that are easily ignored but have better prediction effect. Therefore, The method had better accuracy in predicting cervical lesions.

Keywords: semantic biomedical informatics computing; data mining; high-risk HPV cervical lesion; disease prediction; subgraph mining

check for updates

Citation: Zhou, T.; Xu, P.; Wang, L.; Tang, Y. High-Risk HPV Cervical Lesion Potential Correlations Mining over Large-Scale Knowledge Graphs. *Appl. Sci.* **2024**, *14*, 2456. https://doi.org/10.3390/app14062456

Academic Editors: Artur Jorge Ferreira and Johannes Zenkert

Received: 8 February 2024
Revised: 5 March 2024
Accepted: 12 March 2024
Published: 14 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Patients with high-risk HPV (Human Papillomavirus) infection experience a prolonged state of precancerous lesions (CIN) before their disease progresses to cervical cancer (CC). Cervical cancer is formed when cervical lesions reach grade 3 (CIN3) [1]. Therefore, the best strategy to prevent cervical cancer is to predict the onset of cervical lesions in a timely manner.

There are multiple risk factors that contribute to high-risk HPV infection, and these risk factors also continue to act on the HPV virus to stimulate the expression of its oncogenes, which in turn cause cervical lesions. Therefore, analysis of risk factors is useful in predicting the development of cervical lesions. Some risk factors are directly associated with cervical lesions, while others are indirectly associated due to other diseases or causes. In addition, risk factors have an interactive effect on the development of disease [2]; it does not occur in isolation, and there is an interactive relationship among risk factors. Therefore, it is extremely important to build a knowledge base of risk factors for diseases.

In natural language processing, the construction of ontology knowledge base is usually realized by means of Knowledge Graph. Knowledge Graph describes the objective world entities and their relationships in a structured form, named entity recognition [3] and relationship extraction are Key to Building Knowledge Graphs, and risk factors are also a kind of entities. The identification of entities usually needs to analyze the semantic relationship of the text to achieve, and the relationship between entities can be measured by weights. The size of the weights indicates the closeness of the relationship between the entities.

There are many disease prediction methods in recent years. Statistical methods are widely used in the field of clinical decision. For example, descriptive methods are used to analyze infectious disease problems when the risk factors for the disease are well documented for understanding the variables of interest and their distribution [4]. In addition, machine learning and data mining methods [5] are widely used for disease prediction from case data, which include supervised and unsupervised algorithms. Artificial neural networks can handle various classification problems, and convolutional neural networks in deep learning are also used to extract phenotypes and make risk predictions [6]. Support vector machines [7], decision trees [8], and random forests [9], are also widely used methods. In recent years, these algorithms have been used to structure model and predict risks. Currently, machine learning and data mining methods used to predict disease risk are more accurate than purely statistical methods. A network can be represented as a graph, which consists of nodes and edges. Nodes symbolize entities, while edges symbolize the relationships between entities. The use of network methods is very appropriate when considering the relationship between diseases and symptoms or the synergistic effect of risk factors on diseases [10], which is difficult to predict if considered in isolation. It's applicability to the analysis of genes and phenotypes.

EMR data are often difficult to obtain, with small numbers and many missing values, making it more difficult to predict lesions. Therefore, we must take steps to uncover the key risk factors for disease. In this paper, we constructed the ontology knowledge base of cervical lesions containing risk factors that lead to the occurrence of cervical lesions and the relationships of risk factors by graph structure. Then, we proposed the key risk factor combination mining algorithm, which mines new key risk factors based on common risk factors through a subgraph mining method. In addition, we predicted patient lesions based on the set of key risk factors. The experiments showed that the new key risk factors improved the accuracy of prediction of cervical lesion. The overall design of the lesion prediction method is shown in Figure 1, the details of the knowledge graph construction are shown in the Figure 2.

The contribution of this paper is that it improves the traditional knowledge graph structure by using the word frequency feature of text to measure the closeness of the relationship between entities, which not only quantifies the relationship between entities, but also provides a good foundation for subsequent disease prediction. Secondly, the key node mining algorithm proposed in this paper mines out potential risk factors that have an impact on the disease, which are often overlooked, and provides reference for the prevention and treatment of cervical diseases. In addition, this paper proposes a lesion prediction algorithm based on the set of key risk factors, which improves the accuracy of prediction.

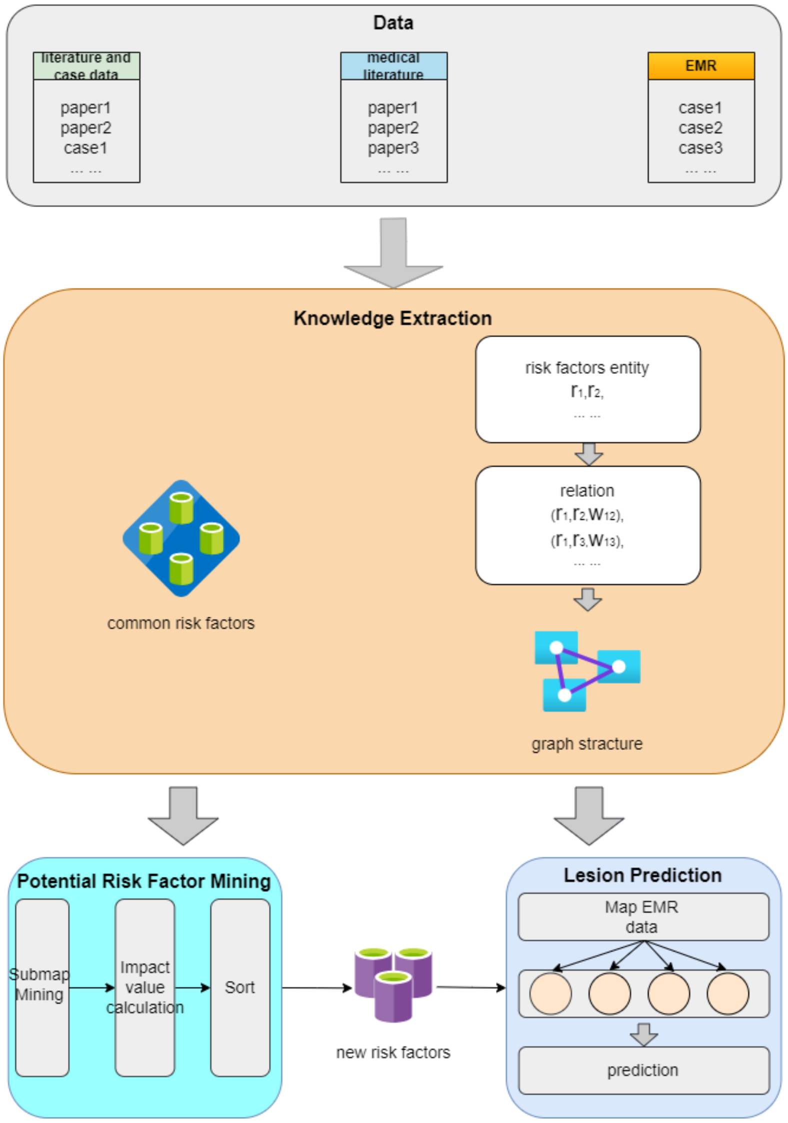

Figure 1. Scheme of the lesion prediction method.

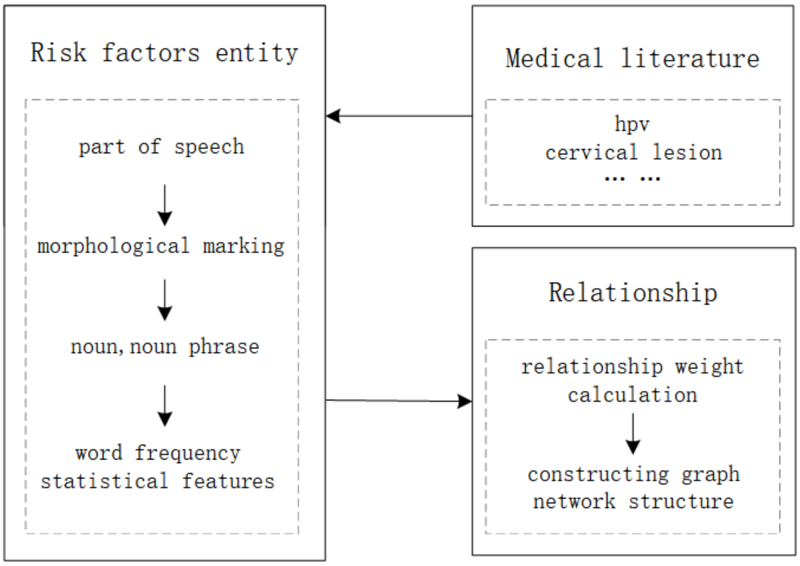

Figure 2. Scheme of knowledge graph construction.

2. Related Work

The construction of knowledge graph has been gradually developed and matured in recent years, and named entity recognition and relationship extraction are the basis of knowledge base construction. Nie, BL. et al. proposed a novel knowledge-enhanced named entity recognition model (KA-NER) [11], which combines the named entity recognition model with the a priori knowledge of the knowledge base, and successfully solves the limitation of the traditional model in utilizing the relevant knowledge. Han, PF et al. established a corresponding entity database based on the statistical collection of Vietnamese names of people, places and organizations in the Vietnamese corpus [12], then, they deeply analyzed the Vietnamese linguistic properties and combined the establishment of the entity database with the proposal of a new named entity recognition model. Zhang, B et al. proposed a deep learning model-based SKG-Learning method [13] for constructing Sentiment Knowledge Graph (SKG) to address the problem of neglecting the relationship between evaluators and evaluation aspects as well as evaluators and topics in traditional sentiment analysis methods. Ji, ZJ et al. provided important theoretical and practical support for the research and application in the field of Chinese knowledge graph construction by designing novel crowdsourcing annotation system, evaluating pre-trained language models [14], and providing open-source datasets and tools to provide important theoretical and practical support for research and application in the field of Chinese knowledge graph construction. Chang, DJ et al. introduced DiaKG, a high-quality Chinese dataset of diabetes knowledge graph, which contains a large amount of entity and relationship information and provides an important resource for the research of domain-specific knowledge graphs in the medical field [15]. Wang, L et al. by proposing a method based on contextual semantic analysis and building a medical knowledge base as well as conducting experimental validation, the paper provides important theoretical and practical support for the tasks of entity recognition and entity relationship extraction in the field of medicine, and provides useful tools and methods for medical knowledge mining and clinical research [16]. Lee, J et al. By pre-trained BioBERT on a large-scale biomedical corpus, using the same architecture as the BERT model [17]. However, BioBERT outperforms the previous BERT model in three typical tasks: biomedical named entity recognition, biomedical relation extraction, and biomedical question answering. The application of knowledge mapping brings us great inspiration to build ontology knowledge base, traditional knowledge mapping has good effect in information retrieval, knowledge management, etc., but if it is to be applied in disease prediction, it needs to be further improved.

Disease prediction methods include statistical methods, machine learning, data mining and deep learning.

Statistical methods for disease prediction are mainly regression analysis and cox risk proportional models. Regression analysis can be divided into linear and nonlinear regression. It identifies quantitative relationships between two or more variables that are dependent on each other. Adithya Mohanavel et al. constructed multiple linear regression models and found that the risk of heart disease increased with smoking and decreased with any form of physical activity [18], effectively making it theoretical. Luo, J et al. proposed a novel approach to analyze the factor group characteristics of the views. Based on logistic regression methods as well as normalized mutual information entropy and information gain rate were used to select the factors. Discriminative minimum class position retention typical correlation analysis was presented and, in addition, a novel model to predict functional risk in the New York Heart Association was proposed [19]. The Cox risk proportional model is also a common regression method used in disease prediction, Zhao, J et al. used multivariate Cox regression analysis and found five non-genetic risk factors associated with the risk of chronic kidney disease [20]. Statistical methods are commonly used by physicians for analysis and are often used for attributional analysis because of their simplicity of thought, but they require more complete information about the case data and have difficulty in collecting and organizing the data.

Compare to purely statistical methods, machine learning, data mining methods and deep learning for predicting disease risk are more accurate [21–23]. Statistical methods have some inherent limitations, correlation tests can be misleading if not designed properly. In addition, statistical methods may give false results if there are any missing variables in the experiment. Farooqui, Md et al. proposed a disease prediction system based on support vector machines and multiple linear regression that can predict possible diseases based on symptoms and it saves the time required for a full diagnosis of the patient [24]. Faruque et al. used several machine learning techniques to explore various risk factors and found that C4.5 decision trees performed better than other algorithms in predicting diabetes, in addition, they identified correlations between different risk factors for diabetes [25], An Y et al. proposed DeepRisk, a fully end-to-end model based on attentional mechanisms and deep neural networks, which not only automatically learns high-quality features from EHRs, but also efficiently integrate heterogeneous and temporally ordered medical data to ultimately predict patients' cardiovascular disease risk [26]. Ahmad Alaiad et al. used a combination of classification and association rule mining techniques to construct an efficient system for predicting and diagnosing chronic kidney disease and its etiology. In addition, the Apriori algorithm was used to discover strong relational rules between attributes, and the application of an integrated approach can significantly improve classification accuracy [27]. Machine learning, data mining and deep learning methods achieve good results in terms of accuracy of prediction, but for medical data, such methods cannot take into account the principles of the disease as well as the relevant factors and are prone to overfitting.

The occurrence of a particular disease is not the result of the independent influence of a factor variable, risk factors have a synergistic effect on the occurrence of the disease, so the use of network graph models can be used to analyze the association between factors and factors and better predict disease. In recent years, the use of graph structure to predict diseases has been gradually applied. EdgCSN was proposed, which is an ensemble learning algorithm that predicts disease genes by means of a network based on clinical samples models trained with centrality features extracted from clinical samples to predict disease genes [28], Fan L et al. proposed a computational framework based on stage-based gene regulatory networks to predict disease genes in breast cancer. Seven stage-based modules were obtained and 20, 12 and 22 key genes were identified for each of the three stages [29]. With the advantage of graph structure, the key influencing factors of the disease can be well analyzed to achieve the effect of predicting the disease.

3. Data Preprocessing

3.1. Text Data Preprocessing

Medical literature is the most authoritative and rigorous textual data containing risk factor knowledge, and the analysis of factors and genes associated with disease occurrence based on medical literature texts is also one of the common tools in bioinformatics. Therefore, mining risk factors and conducting factor relationship analysis based on medical literature is an effective means. Case data are in various forms with many missing values and also require preprocessing. Therefore, the data preprocessing section contains preprocessing of the medical literature text as well as preprocessing of the case data.

The preprocessing steps of medical literatures are divided into storage type conversion, meaningless component cleaning, word separation and part-of-speech tagging. The collected literature data are all in pdf storage format, we used conversion software to convert the pdf format to word format, and then stored as a txt file with paragraphs as the unit. Therefore, medical literature contains a large number of meaningless components, such as pictures, links, names of people, etc. If all of them are retained, a large amount of computing resources will be wasted. Therefore, We cleaned the meaningless components by observing the characteristics of the text dataset and writing regular expressions for extracting plain text to achieve the removal of invalid information and facilitate subsequent processing and analysis. The cleaned text data is divided into words, NLTK package is fast and high quality, we also uses nltk to divide the text into words. The lexical annotation work is carried out directly after word separation, which can improve the accuracy of lexical annotation. Deactivated words are words that have no value to the semantic expression of the text but appear more frequently, such as "he", "that", "about" and so on. The presence of deactivated words in the text not only interferes with valuable information, but also causes excessive content in the text and wastes computational resources. In this part, we create a deactivation word list, match the words in the lexicon with the deactivation word list, delete all the words that appear in the deactivation word list, and keep all the remaining text to realize filtering the words in the lexicon.

EMR data includes electronic case data in text form and data collected in tabular form. The Electronic Medical Record (EMR) data in text form are preprocessed as follows: scanning the cases, converting the storage format, converting the text in various formats to English text format; extracting nouns and noun phrases by word division and lexical annotation; constructing a negative word list based on the text characteristics, and using the sentences where the negative words are located to remove key words that are not abnormal in the cases, such as "deny hypertension, diabetes, coronary heart disease", etc. Finally, we constructed a set of key words for the case. The set of case key words was constructed. The data in the form of a table is preprocessed as follows: attributes are risk factors, and values are the values of their risk factor profiles. The value types include continuous values and binary values. For example, "duration of smoking (years): 2" means two years of smoking, and "immunodeficiency virus: 1" means the patient is infected with immunodeficiency virus, while a value of 0 means no infection. The preprocessing of this dataset is mainly for missing values. If 80 percent of the data of this case are missing values then they will be deleted and in addition the missing values will be uniformly represented by 0.

3.2. Risk Factor Background

Although human papillomavirus infection is a major risk factor for the development of cervical lesions, there is growing evidence that multiple environmental factors influence the development of cervical lesions and the clinical course of cervical lesions. Multiple risk factors have a synergistic effect on disease development, for example, one study showed that the superimposed effect of hormonal contraceptive use, smoking habits and HPV infection was higher than the risk of both hormonal contraceptives and HPV. That implies that there is a relationship between risk factors in the development of cervical cancer. When two known factors, A and C, are linked to the occurrence of a disease, and another factor, B, is associated with both A and C, then B also plays a crucial role in how A and C contribute to the occurrence of the disease. Moreover, B might be a previously unnoticed factor, so it's necessary for us to analyze its role.

The association of risk factors can be represented by the graphical structure, In graph structure, nodes can be connected by edges if they are related to each other. Two nodes may be directly connected to each other or there may be multiple paths through other points. Since many people like to focus on the shortest path between two points, the points on these paths are often ignored, however, these nodes are not only closely related to them but also may have undiscovered knowledge. In the risk factor network structure, finding the implied nodes on the path between two common risk factor nodes and analyzing them can help us to uncover new potential risk factors, which is of crucial importance for the study of cervical lesions. As shown in Figure 3a, v1,v2 are two common risk factor nodes, and there are three paths v1-v2, v1-v3-v2, v1-v2-v5-v2 between them. v1 and v2 are directly connected, which indicates that there is a relationship between them, in addition, they are also indirectly related through v3, v4, v5, then these three nodes are important for us to uncover new risk factors is of great significance, as shown in Figure 3b, there may be new knowledge in these nodes that also have an impact on the disease and have not been discovered yet, so, through graph structure analysis, by mining new risk factor nodes among common nodes, it can help us to have an updated understanding of cervical lesions and help to improve the accuracy of lesion prediction.

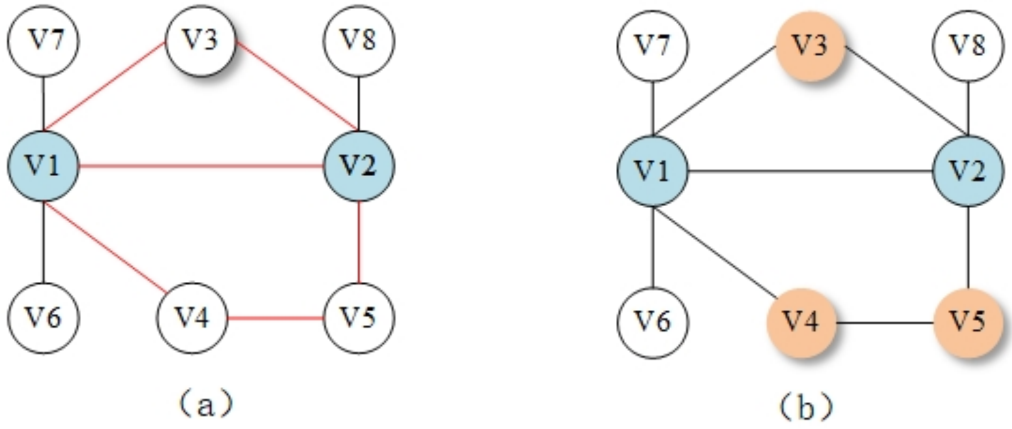

Figure 3. Risk factor graph mining example: (**a**) is original dataset graph, V1 and V2 are basic risk factors. After caclulating the correlations (paths) among the whole large-scale graphs, V3, V4, and V5 are newly mined risk factors by subgraph mining method, as shown in (**b**).

Usually, electronic medical records are in the form of text, in order to perform disease prediction, most of the current methods are by manually transforming the case text into a representation of attributes and values. However, the case text does not show all the attributes, which leads to the occurrence of a lot of missing values, which has a large impact on the prediction results. I propose a disease prediction method, borrowing the constructed graph network, by directly extracting the keywords in the case text can carry out disease prediction, which can avoid the bias of the disease prediction caused by more missing values.

4. Ontology Knowledge Base Construction Based on Semantic Relationship

Risk factors have an interactive effect on the development of disease, which does not occur in isolation, and there are interactions among risk factors. The construction of an ontology knowledge base for cervical cancer can provide a clear understanding of the causative risk factors and discover the relationships among them, which is important for the mining of key risk factors and disease prediction.

Risk factors are usually expressed as noun words and phrases, we propose a Textual Risk Factor Extraction Method Based on Lexical and Grammatical Patterns (TRFLEX-LGP) model, so the text can be extracted based on the lexicality after lexical annotation. Word extraction includes extracting words of "noun", "noun plural", "proper noun" and "proper noun plural words of "noun", "noun plural", "proper noun" and "proper noun plural". The extraction of phrases is based on the grammatical pattern of noun phrases, specifically: "adjective and noun word and preposition and adjective and noun word", "adjective and noun word", "noun word and noun words". The noun words include "noun", "noun plural", "proper noun" and "proper noun plural". All the extracted words are de-duplicated, and finally the de-duplicated words are stored in the thesaurus.

Word frequency has always been a feature of the text to measure the importance of keywords. Similarly, the degree of relationship between two keywords can be defined in terms of the frequency of simultaneous occurrences of two words. Therefore, the keywords in the keyword lexicon are next formed into multiple phrases of length 2. The frequency of each pair of words in all sentences, paragraphs and articles is counted separately to form a collection of word frequency and stored in the database.

The risk factor knowledge base is constructed using a relational graph network structure because the network graph structure is able to reflect the relationships of factors. The network graph structure is composed of nodes, edges and the weights of the edges. The risk factors are the nodes in the graph, and the risk factors can be connected with edges if there is a relationship between them. The weights of the edges reflect the strength of the relationship between the risk factors. In this section, the network structure is constructed using a graph with the right undirected graph.

The noun words and phrases were extracted and added to the lexicon in the data preprocessing stage, and all the words in the lexicon were made as nodes in the network structure of the risk factor relationship graph in this part, forming the set of nodes N. The frequency of two keywords co-occurring in the same paragraph in the medical literature indicates the existence of a relationship between these two keywords, so this part filters the word frequency statistics of the keyword phrases in the database, screen the phrases whose frequency is not 0, and associate the nodes corresponding to these phrases in the graph network structure to form the set of edges E.

Parameters' notations for our proposed methods are shown in Table 1.

Table 1. Notations.

Notation	Description
K	Keyword dataset.
KC	The keyword binary phrase set of K.
W_{ij}	The strength of the relationship of keyword i and keyword j.
A_{ij}	Keyword i co-occurring with keyword j in the same article.
P_{ij}	Keyword i co-occurring with keyword j in the same paragraph.
S_{ij}	Keyword i co-occurring with keyword j in the same sentence.
SP_i	The sum of the word frequencies of i and all keywords co-occurring with i in the same paragraph.
SS_i	The sum of the word frequencies of i and all keywords co-occurring with i in the same sentence.
SA_i	The sum of the word frequencies of i and all keywords co-occurring with i in the same article.
SN	The seed node set.
SNC	The keyword nodes in it are combined two by two to form the seed binary set.
EF_i	The influence strength of keyword i.
BC_i	The mesoscopic centrality of keyword i in the current subgraph.
C_i	The number of occurrences of keyword i in all subgraphs.
$Ctopi$	The number of occurrences of keyword i in top.
P	The threshold value.
F	The set of key risk factors for cervical lesion.
H	The set of case keywords.
top	The set of keyword nodes in each subgraph whose node mesoscopic centrality value is greater than the three-fourths quantile of the mesoscopic centrality of all nodes in the current subgraph.
N	The set of nodes in risk factors keyword relationship network graph.
E	The set of edges in risk factors keyword relationship network graph.
G	The risk factors keyword relationship network graph.
f_y	Any one in F.
h_x	Any one keyword in the set of case keywords H.
$Z_{(hx,fy)}$	The value of the relationship between h_x and f_y.
num_y	Relations between the set of case keywords H and f_y.
map_y	Mapping value of the case keyword set H to the key risk factors.
$flevel$	Patient risk level.

The strength of the relationship of keyword i and keyword j W_{ij}:

$$W_{ij} = exp\left(\frac{A_{ij}}{SA_i + SA_j - A_{ij}}\right) + exp\left(\frac{P_{ij}}{SP_i + SP_j - P_{ij}}\right) + exp\left(\frac{S_{ij}}{SS_i + SS_j - S_{ij}}\right) \quad (1)$$

The keywords in the keyword dataset K are combined two by two to form the keyword binary phrase set K, $(i,j) \in KC$. The frequency A_{ij} is keyword i co-occurring with keyword j in the same article; the frequency P_{ij} is counts of co-occurring in the same paragraph; the frequency S_{ij} is counts of co-occurring in the same sentence. SA_i denotes the sum of the word frequencies of i and all keywords co-occurring with i in the same article. SP_i denotes the sum of the word frequencies of i and all keywords co-occurring with i in the same paragraph. SS_i denotes the sum of the word frequencies of i and all keywords co-occurring with i in the same sentence.

The keyword word frequency statistics greater than 2 in K is represented as the set of nodes N in the graph, and the connection between keywords is represented as the set of edges E in the graph, which denotes the weights of edges between node i and node j, W is the set of all edge weights. The keyword relationship network graph model is constructed as $G = (N,E,W)$.

We have further improved the traditional knowledge graph approach by utilizing the word frequency features of the entities and calculating the specific weights to present the relationships between the entities.

5. Key Risk Factor Combination Mining Algorithm

Subgraph is one of the basic concepts of graph theory, which has node set and edge set as the graph of a subset of node set and subset of edge set of a certain graph. The ontology knowledge base is a huge relational graph structure, subgraph mining method could improve computational efficiency. We proposed Key Risk Factor Combination (KRFC) mining algorithm, which used common risk factors of cervical cancer as seed nodes and mined new key risk factors using subgraph mining method. The algorithm idea is shown in Figure 4.

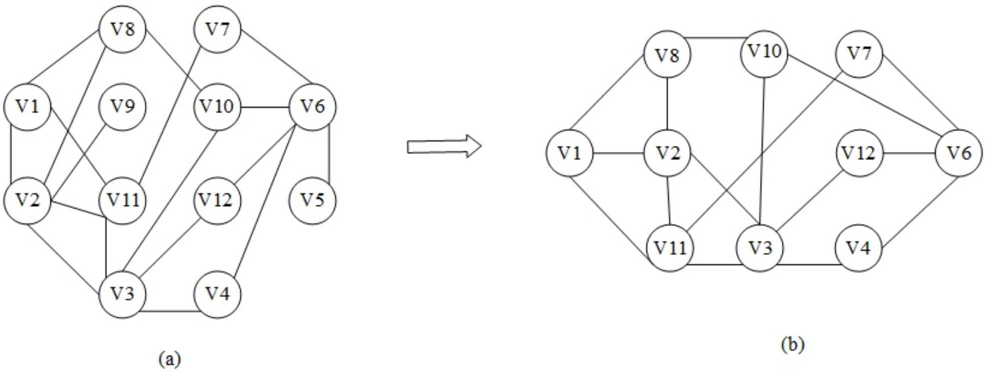

Figure 4. Combined key risk factors mining example: (**a**) is a complex original graph, and (**b**) is a mined strong association graph after depth-first searching based on risk factors clusters.

Figure 4a represents the risk factor relationship network structure, v1–v12 are keyword nodes, this algorithm combines common key risk factors two by two, takes any one risk factor in the combination as the starting point and depth-first search, finds another node within the minimum consumption and forms a subgraph Figure 4b with all nodes and edges on the path, and then performs further impact value analysis and calculation on all nodes in the subgraph to get the new key risk factor.

The algorithm flow of Key Risk Factor Combination Mining Algorithm is as follows: first, the common risk factors are made as seed node sets, and in the risk factor graph structure, starting from a seed node, traverse the neighboring nodes using depth-first search until another node is found, which forms a subgraph containing two seed nodes and other nodes on the path between them. nodes and other node subgraphs on the path between them. After traversing all the nodes, some new key risk factors can be obtained by extracting features for all the other nodes in the subgraph, calculating the influence intensity of the nodes, and setting a threshold. Combining the new key risk factors with the seed nodes constitutes the full set of key risk factors for the disease.

Based on the seed node set SN, the keyword nodes in it are combined two by two to form the seed binary set SNC, m, n are any two seed nodes in SN and,define Cost as the minimum consumption to find another target node from a point in the graph. Using m

as the starting point, the $1 - hop$ node is searched based on the $top - N(N = 100)$ of the weights of the connected edges of the starting node, and the strongly connected subgraph containing a specific node is generated by searching for another kind of subnode n in Cost. Where Cost = 5000 nodes, the threshold can be adjusted according to the actual situation.

Our proposed KRFC algorithm pseudo-code is shown in Algorithm 1.

Algorithm 1 KRFC algorithm.

```
Input: G, common risk factors set
Output: new set of key risk factors
SNC ← paired combination of SN
SUB ← empty set
for each $(i, j)$ in SNC do
    Cost ← Number of nodes needed to find $j$ from $i$
    if Cost $< 5000$ then
        $Sub_{ij}$ ← subgraph of G containing $i$ and $j$
    else
        Continue
    end if
    SUB append $Sub_{ij}$
    for each node $n$ in $Sub_{ij}$ do
        $P_{ij}^{n}$ ← Number of shortest paths connecting $i$ and $j$ and passing through node $n$
        $Q_{ij}$ ← Number of shortest paths connecting $i$ and $j$
        $BC_n$ ← sum$(\frac{P_{ij}^{n}}{Q_{ij}})$, $i \neq n \neq j$ (mesocentricity of $n$)
        top ← the set of nodes in $Sub_{ij}$ whose Satisfied $BC_n$
    end for
end for
for each $n$ in SUB do
    F ← empty set
    $C_n$ ← the number of occurrences of $n$ in SUB
    $C_{top_n}$ ← the number of occurrences of $n$ in top
    $EF_n$ ← Influence value $(BC_n, C_n, C_{\text{top}_n}, \varepsilon)$
    if $EF_n >$ threshold value $\rho$ then
        $F \leftarrow SN$ append $n$
    else
        Continue
    end if
    return $F$
end for
```

The influence strength of keyword i EF:

$$EF_i = \frac{\sum_{n=1}^{n^*} BC_i * \frac{C_i}{n^*}}{\frac{Ctop_i}{C_i}} \quad (\frac{C_i}{n^*} > \varepsilon) \tag{2}$$

n^* indicates the number of subgraphs generated by all seed nodes, BC_i is the mesoscopic centrality of keyword i in the current subgraph, C_i indicates the number of occurrences of keyword i in all subgraphs. top indicates the set of keyword nodes in each subgraph whose node mesoscopic centrality value is greater than the three-fourths quantile of the mesoscopic centrality of all nodes in the current subgraph. $Ctop_i$ denotes the number of occurrences of keyword i in top.

Set the threshold value $\varepsilon = 0.01$, $\frac{C_i}{n^*} > \varepsilon$, mean denotes the mean value of the mediocentricity of all keyword nodes in top, and std denotes the standard deviation of the mediocentricity of all keyword nodes in top.

The threshold value ρ:

$$\rho = mean + \theta * std, \quad \theta = 0.6 \sum_{n=1}^{n^*} BC_i > \rho \tag{3}$$

It can be adjusted ε and θ according to the actual situation.

The impact intensity of all keywords that meet the threshold setting is calculated to obtain the $top - k$ impact intensity, and the keywords corresponding to the $top - k$ impact intensity constitute the set of key risk factors for cervical lesion $F = \{f_1, f_2, f_3, \ldots, f_v\}$.

6. Lesion Prediction Algorithm Base on Mapping

There are many missing values in case data, and traditional missing value processing method is not suitable for case data, because any filling method with missing values may cause deviation in prediction results. Therefore, this paper proposes a disease prediction algorithm based on mapping principle, which maps the risk factor keywords extracted from EMR data to key risk factor nodes in the graph model structure. To obtain different relationship values, and in using the different relationship values for lesion likelihood prediction. The idea of lesion prediction base on mapping (LPM) algorithm is shown in Figure 5.

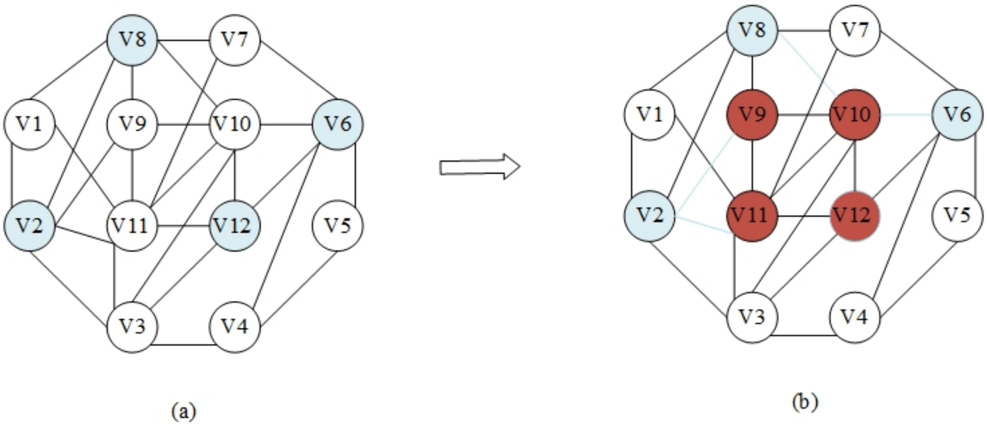

Figure 5. Lesion prediction process example: V2,V6,V8, and V12 are the basic risk factors from patient EMR in (**a**), and V9,V10,V11,and V12 are the newly extracted and confirm the key risk factors based on potential relationship mining, as shown in (**b**).

As shown in Figure 5a, {v2,v6,v8,v12} denote the set of keywords extracted from one patient case data, they exist in the risk factor relationship graph network structure, As shown in Figure 5b, {v9,v10,v11,v12} denote all the key risk factors we get, the keywords in the set of keywords extracted from patient case data may exist in the set of key risk factors or beyond it, therefore, we design the LPM algorithm to map the set of keywords extracted from each patient case data to the key risk factors uniformly, which can get different case information features.

Our proposed LPM algorithm pseudo-code is shown in Algorithm 2.

Algorithm 2 LPM algorithm.

```
Input: G, $F$, EMR data
Output: Lesion risk level
$H \leftarrow$ extract abnormal keywords of EMR data
for each $h_x$ in $H$ do
    for $f_y$ in $F$ do
        $Z_{(hx,fy)} \leftarrow$ the value of the relationship between $h_x$ and $f_y$
        if $h_x = f_y$ then
            $Z_{(hx,fy)} = \max$(neighbourhood weight of $f_y$ in G)
        else
            if directly connected between $h_x$ and $f_y$ then
                $Z_{(hx,fy)} = W_{hx,fy}$ in G
            else
                if presence path between $h_x$ and $f_y$ then
                    $mul \leftarrow$ the concatenated product on path
                    $npath \leftarrow$ path length
                    $Z_{(hx,fy)} = \frac{mul}{npath}$
                else
                    $Z_{(hx,fy)} = 0$
                end if
            end if
        end if
        $map_{fy} \leftarrow$ mapvalue($Z_{(hx,fy)}$)
    end for
    $MAP = (map_1, map_2, map_3, ...map_v)^T$
end for
$flevel \leftarrow$ sigmoid($MAP$)
```

h_x is any one keyword in the set of case keywords H, $x \in (1, u)$, f_y is any one in F, $y \in (1, v)$, $Z_{(hx,fy)}$ is the value of the relationship between h_x and f_y.

When $h_x = f_y$,$Z_{(hx,fy)}$ is the maximum value of the edges connected to all neighboring nodes of f_y in the graph model G. When $h_x \neq f_y$, and there is a directly connected edge between h_x and f_y in the graph model G, $Z_{(hx,fy)} = Z_{(hx,fy)}$. When $h_x \neq f_y$, and there is no directly connected edge between h_x and f_y in the graph model G, but there is a path, define the concatenated product of the strength values of the relations passing on the shortest path as mul, and the path length as $npath$. $Z_{(hx,fy)}$ was calculated using Equation (4). When $h_x \neq f_y$, and there is no directly connected edge between h_x and f_y in the graph model G, but also not exsit the path, $Z_{(hx,fy)} = 0$.

$$Z_{(h_x,f_y)} = \frac{mul}{npath} \tag{4}$$

the statistical number num_y of relations between the set of case keywords H and f_y:

$$num_y = \sum_{x=1}^{u} 1 \; Z_{(h_x,f_y)} \neq 0 \tag{5}$$

The mapping value of the case keyword set H to the key risk factors:

$$map_y = \frac{\sum\limits_{x=1}^{u} Z_{(h_x,f_y)}}{num_y} \tag{6}$$

The mapping matrix $MAP = (map_1, map_2, map_3, ..., map_v)^T$ of cases is obtained by calculating the mapping values of all key risk factors. defined a^T. b is parameter values in the *sigmod* function obtained by logistic regression model training.

MAP is used as the variable value of the case, and the patient risk level $flevel$ is calculated using the $sigmod(MAP, \alpha^T, b)$. $flevel >= 0.6$, patient with high risk level and high possibility of lesion. $0.6 > flevel >= 0.4$, patient with medium risk level and medium possibility of lesion. $flevel < 0.4$, patient with low risk level and low possibility of lesion.

7. Results

The datasets of this paper are medical literature dataset and EMR dataset. The medical literature was collected by randomly downloading a total of 2221 relevant articles on the PubMed website with cervical lesions as the keyword as the medical literature dataset. The Electronic Health Record data were collected from the following ways: hospital, case report articles of cervical lesions on the PubMed website [30], and clinical data on cervical cancer (CESC) from The Cancer Genome Atlas (TCGA) database [31]. A total of 371 case data were collected, a total of 37 electronic health records, 27 case reports downloaded from PubMed and 307 cases of TCGA-CESC. Positive sample 307 cases, negative sample 64 cases. The detailed scale and presentation of the data in this study is shown in Tables 2–4, and the experimental parameter settings are shown in Table 5.

Table 2. Table of text sizes.

Articles	Paragraphs	Word	Bytes
2221	999,445	14,942,952	1,685,262,336

Table 3. Table of graph scale.

Nodes	Edges	Average Degree
116,336	18,013,237	310

Table 4. Table of case sizes.

Case Dataset	Total Count	Positive Sample Count	Negative Sample Count
All case data	371	307	64
TCGA-CESC	307	307	0
EMR	37	0	37
case report	27	0	27

Table 5. Table of experiment parameters.

Cost	ϵ	θ	Cross-Validation Folds
5000	0.01	0.6	5

The experimental environment is Windows 10 operating system with Python3.7 and Networkx2.1.

7.1. Ontology Knowledge Base

Before forming the final graph structure, we filtered the extracted words and phrases several times in order to remove useless words and improve the computational efficiency of the graph structure. We collected 45 common risk factors for cervical cancer, and matched these 45 risk factors proximally in phrases extracted through medical literature texts to obtain a knowledge base of 65,536 phrases about cervical cancer risk factors, which was used as a standard library to run the WBC and BioBERT algorithms. Finally, our constructed the large-scale knowledge graph contains 116,336 nodes, which contain 50,800 words and 65,536 high-quality phrases.

In order to test our proposed TRFLEX-LGP method for high-quality phrases extraction performance, we compare with two large-scale language processing models WBC [16] and BioBERT [17], the results as shown in Table 6. In the testing, our proposed TRFLEX-LGP model is better than BioBERT. WBC model actually is our newly proposed high-quality phrase extraction model in our lab, which is focus on solving long phrase, rare words and professional phrase recognition tool for medical texts. TRFLEX-LGP model is closed to WBC model about high-quality phrases extraction, but our proposed TRFLEX-LGP is focus on calculating and mining the correlations among extracted phrases, not only for extracting the phrases. So, TRFLEX-LGP is better for extracting the high-quality phrases more related to seed nodes based on knowledge base.

Table 6. Comparison of high-quality phrase rates.

Methods	Extracted Phrases Number	High-Quality Phrase Recognition Rate
WBC	101,496	64.57%
BioBERT	103,462	63.34%
TRFLEX-LGP	102,898	63.69%

The graph structure G is constructed by semantic relation computation. G stored into the Neo4j graphical database. The whole risk factor network structure covers more than 100,000 keyword nodes and more than 10 million relationships between keywords, and some of the node visualization results are shown in Figure 6.

The degree distribution represents the distribution of connected edges in the network. The degree of a node is defined as the number of edges directly connected to that node. The power law distribution is often referred to as the Matthew effect and the law of two or eight. The power-law distribution characteristic of degree has a great impact on the fault tolerance and aggressiveness of the network. Therefore, we analyze the power law distribution of the constructed network. The power law distribution of the degree of the risk factor relationship graph network structure is shown in Figure 7. The degrees of the nodes in the graph structure constructed in this paper conform to a power law distribution.

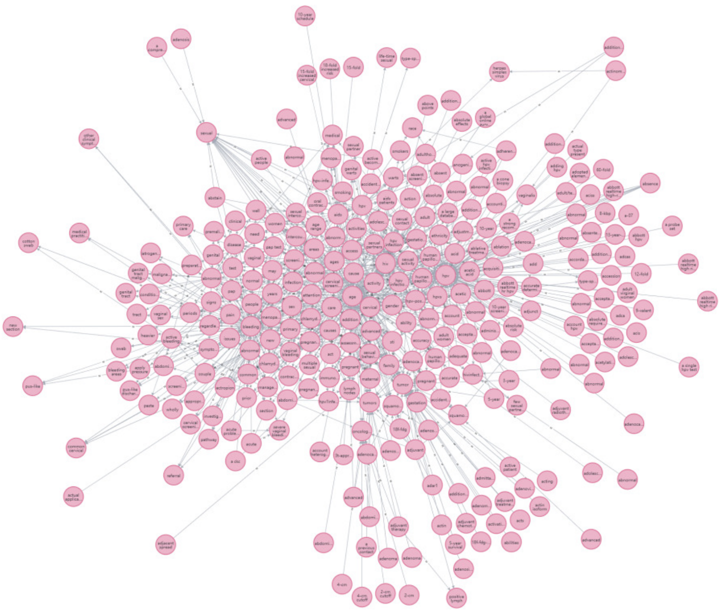

Figure 6. Sample nodes visualization of constructed HPV related medical entities large-scale graph.

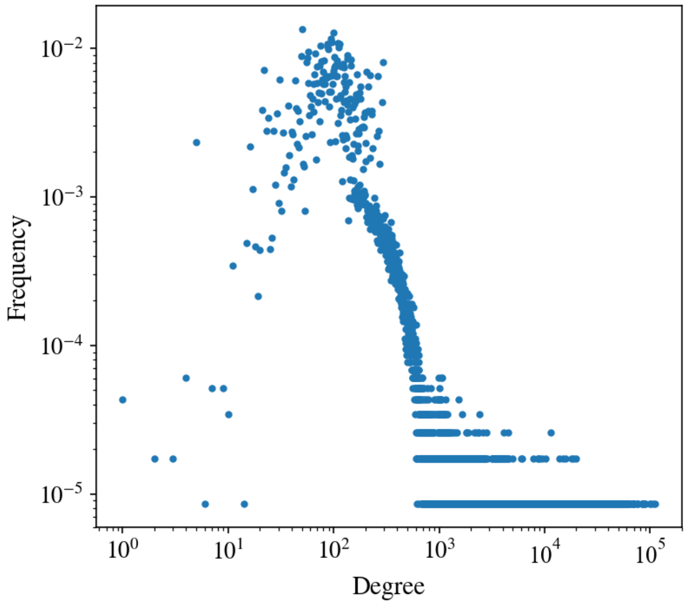

Figure 7. Power-law distribution of network structure degree of risk factor relationship diagram.

7.2. Key Risk Factors Combination Mining

Using the cervical cancer risk factor guidelines [32] and the UCI case dataset [33], we compiled the common risk factors as shown in Table 7. Using the common risk factors as seed nodes, new risk factors were mined based on the key risk factor combination mining algorithm as shown in Table 8. We found that anxiety and depression were also important factors influencing disease progression, in addition to the vaginal microenvironment as well as hormones that contribute to HPV infection and promote the development of lesions. Persistent HPV infection could greatly contribute to the development of lesions, and this should be taken into account in the pathological diagnosis.

The KSGC algorithm [34] is selected as the comparison algorithm in this experiment. The KSGC is a gravity formula based critical node identification algorithm that combines degree centrality and K-shell to measure the importance of a node in the propagation dynamics. The KRFC algorithm proposed in this paper is based on meso centrality and importance of nodes in paths as well as subgraphs to find critical nodes. As shown in Table 9, the nodes mined by the KRFC algorithm proposed in this topic are compared with the KSGC algorithm, and the same calculation method is chosen for the threshold selection, Table 9 shows the risk factor nodes of the $top - 20$ node ranking of the mined nodes. From the table, we can observe that the $top - 20$ nodes mined by the KSGC algorithm are all common risk factors, i.e., the seed node set in the KRFC algorithm of this topic, while the $top - 20$ nodes mined by the KRFC algorithm contain not only some common risk factors, but also many other risk factor keywords such as “hormonal The $top - 20$ nodes of KRFC algorithm not only contain some common risk factors, but also many other risk factors such as “hormonal”, “estrogen”, “anxiety”, etc. These risk factors have important reference significance for disease research and can provide new perspectives for clinical research. Therefore, the comparison reveals that the algorithm proposed in this topic is more suitable for mining potential risk factors, which is important for disease clinical research.

Table 7. Common risk factors.

Common Risk Factors	Common Risk Factors
age	intrauterine device (IUD)
Alcohol	Sexually transmitted diseases
Drinking time	Condyloma
Amount of alcohol consumption	Syphilis
Smoking	Genital Herpes
Smoking time	AIDS
Amount of smoking	HIV
Abnormal age of menarche	Herpes simplex virus
Age of menopause	Chlamydia
Young age of first intercourse	Chlamydia trachomatis
Sexual partners	Hepatitis B
High number of sexual partners	Warts
Sexual behavior	Vaginitis
Oral contraceptives	Pelvic inflammatory disease
Pregnancy	Immunosuppression
Young age at first full-term pregnancy	Lymph node abnormalities
Normal birth	Postmenopausal bleeding
Multiple pregnancy	Vaginal bleeding
Miscarriage	Bleeding after sexual intercourse
Ectopic pregnancy	Abdominal pain
Stillbirth	Other tumors
Human papillomavirus types	Family history of disease
Diethylstilbestrol (DES)	

Table 8. New risk factors.

New Risk Factors	Keywords with High Relationship Intensity in the Graph
Vaginal discharge	vaginal bleeding, Bleeding after sexual intercourse, precancerous lesion, cervical HPV infection
hormone	estrogen, progesterone, pregnancy, antibiotics, HPV16 e7
estrogen	Hormone, progesterone, pregnant women, hr-HPV, infections, perinatal
progesterone	pregnancy, hrHPV infection, immunity
Vaginal microbe	lactobacillus, vaginal PH, acids, genital infection
lactobacillus	Vaginal microbe, vaginal PH, Inflammatory diseases, HPV infections
anxiety	depression, cellular immunity, HPV-infected cells
depression	anxiety, immunity, HPV infect
Persistent hpv	HPV, early stage cervical cancer, hsil, lsil

Table 9. Comparison results of KSGC and KRFC.

KSGC	KRFC
tumor	hpv
pregnancy	tobacco
warts	immunosuppression
hpv	hormonal
smoke	tumor
hiv	sexual intercourse
immunosuppression	pregnancy
syphilis	vaginalis
miscarriage	estrogen
menopause	persistent infection

Table 9. *Cont.*

KSGC	KRFC
age	postmenopausal
hepatitis	aids
herpes	age
chlamydia	immune system
lymph	vaginal discharge
condyloma	lymph nodes
stillbirths	chlamydia trachomatis
postmenopausal	follow-up
aids	anxiety
follow-up	herpes simplex virus

Before calculating the node impact values, we used three features to describe the keyword nodes, namely BC_i, C_i and $Ctop_i$, for the newly mined key risk factors, we analyzed these three features for them, as shown in Figures 8–10.

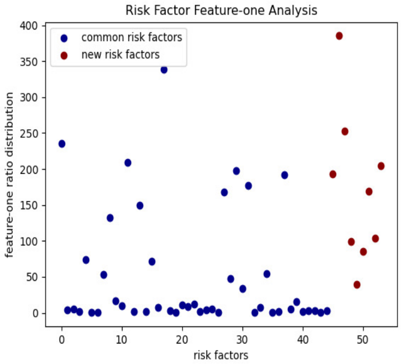

Figure 8. Risk factor node feature BC_i ratio distribution analysis.

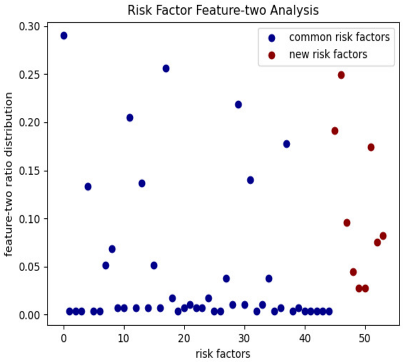

Figure 9. Risk factor node feature C_i ratio distribution analysis.

Figure 8 shows the values of the nodes, and the median of all nodes is selected as the benchmark to view the ratio distribution of the values of the key risk factors, blue represents the seed risk factors, and red represents the newly mined risk factors, which shows that the newly mined factors occupy a relatively very important position. As shown in Figure 9, higher values indicate that the nodes appear more frequently in all mined subgraphs, as seen in the high frequency of newly mined nodes, which is also able to determine the importance of new key risk factors.

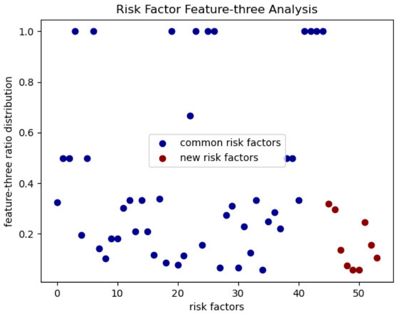

Figure 10. Risk factor node feature $Ctop_i$ ratio distribution analysis.

As shown in Figure 10, higher $Ctop_i$ indicates that nodes appear in top and rarely outside top, while lower $Ctop_i$ indicates that nodes not only appear in top but also appear frequently outside top, which is sufficient to demonstrate the importance of the role played by nodes in the knowledge base.

7.3. Lesion Prediction

The correlation analysis of the mapped attribute matrix is shown in the Figure 11.

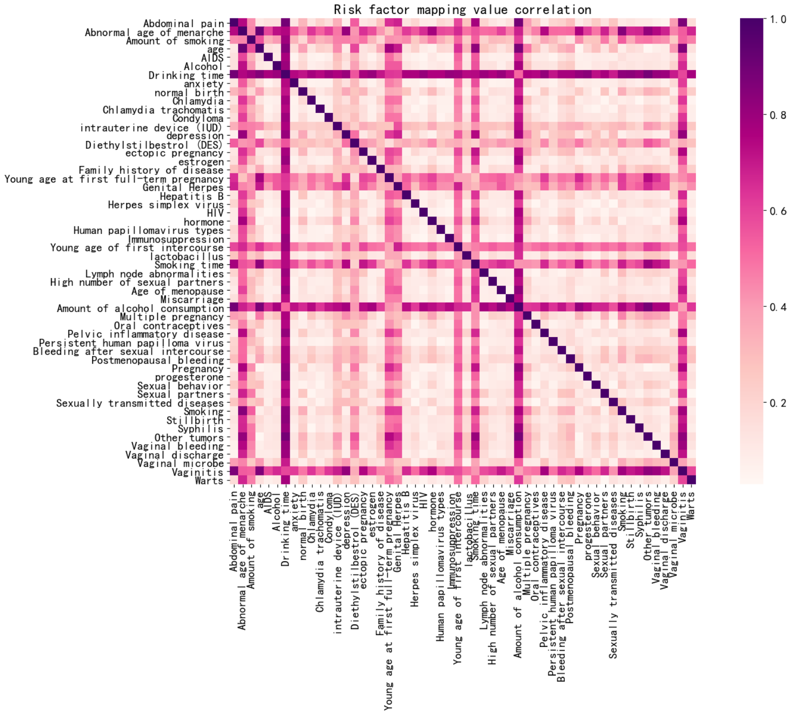

Figure 11. Correlation analysis of risk factor mapping values.

The evaluation indicators for this experiment are as follows:

TP (True Positive): True case, both true and predicted values are positive cases.

FP (False Positive): False Positive, where the true value is negative and the predicted value is positive.

FN (False Negative): False negative case, the true value is positive, the predicted value is negative.

TN (True Negative): True negative case, both the true value and the predicted value are negative cases.

True positive rate: $TPR = \frac{TP}{TP+FN}$

False positive rate: $FPR = \frac{FP}{FP+TN}$

Precision: $P = \frac{TP}{TP+FP}$

Recall: $R = \frac{TP}{TP+FN}$

$F1$ score: $F1$ score $= \frac{2\times P\times R}{P+R}$

The horizontal coordinate of the ROC curve is FPR and the vertical coordinate is TPR, the prediction results are sorted according to the predicted positive class probability value, the threshold value is gradually reduced from 1, and the samples are predicted as positive cases one by one in this order, and the current FPR and TPR values can be calculated each time, and the images are plotted with TPR as the vertical coordinate and FPR as the horizontal coordinate.

In this paper, we take the 5-fold cross-validation method to train the data and select lasso logistic regression [35] and Svm-AdaBoost [36] as comparison experiments, The precision, recall, and F1score comparisons of the three algorithms are shown in Figure 12 and Table 10, where the overall performance of our algorithm is better.

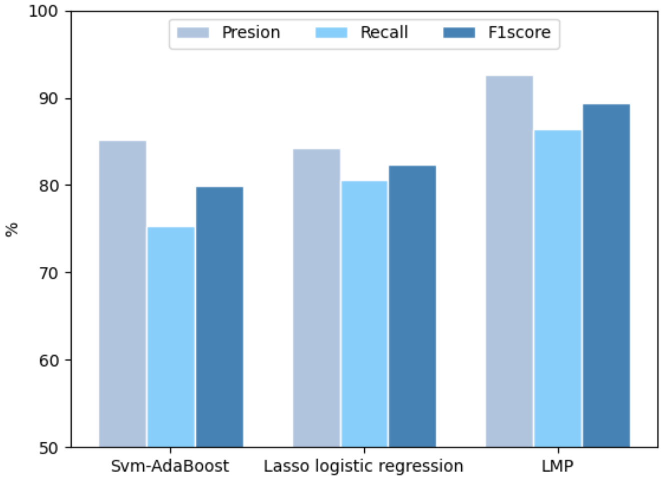

Figure 12. Comparative Experimental Analysis.

Table 10. Comparative experimental data analysis.

	Precision	**Recall**	**F1 Score**
Svm-AdaBoost	85.21%	75.38%	79.99%
Lasso logistic regression	84.25%	80.60%	82.38%
LMP	92.59%	86.47%	89.43%

The *Cost* parameter in this study is crucial, as it represents the minimum expense for one node to reach another. In order to explore the optimal value of *Cost*, we experimented with different *Cost* values and observed the F1 score performance of the method, as shown in Figure 13. We found that as the *Cost* increases to 5000, the performance gradually stabilizes, which is why we chose a *Cost* of 5000.

Each time a result is plotted a ROC curve and the area under the curve is calculated, as shown in Figures 14–16, the blue curve is the average ROC curve, and in Figure 17, we compare the average ROC curve and the area under the curve of the three algorithms.

As shown in Figures 14–16, our algorithm is more stable compared to the other two algorithms, with a small range of fluctuation in the area under the curve for each fold. In addition, observing Figure 17, the average ROC curve of the LMP algorithm is closer to the upper left corner, and the area under the curve is also the maximum in comparison, which indicates that our algorithm has good results in terms of accuracy as well as stability of the model.

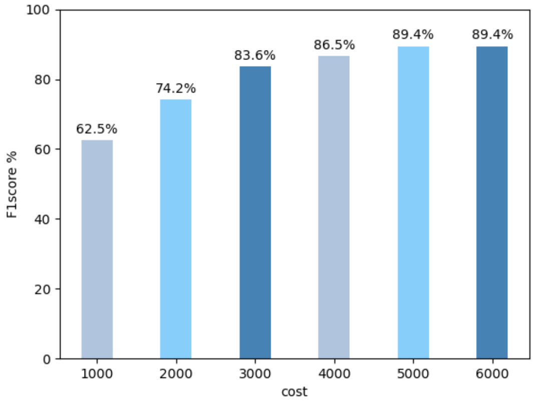

Figure 13. Change in F1 score for different *Cost* values.

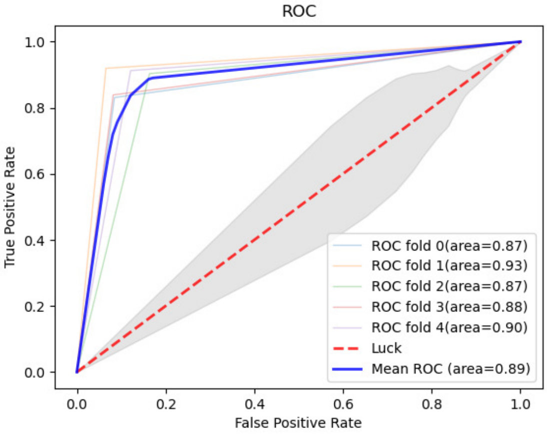

Figure 14. LMP algorithm 5- fold cross validation roc curve.

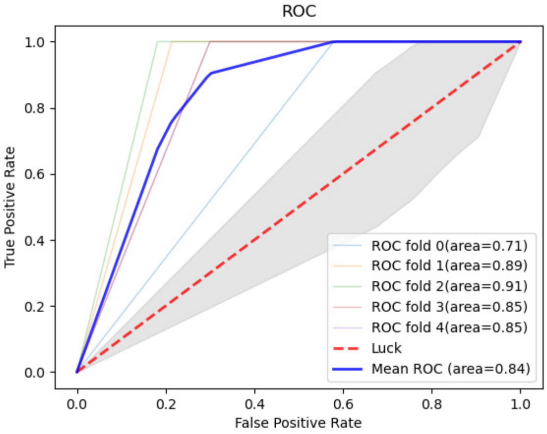

Figure 15. Lasso logistic regression 5- fold cross validation roc curve.

The above three datasets were validated and compared separately using 5-fold cross-validation for further validate the effectiveness of the algorithm in this study, the accuracy of each 5-fold was averaged for this algorithm accuracy. As shown in Figure 18, the accuracy of our proposed LMP algorithm was significantly higher than the other two comparison algorithms for the case report dataset collected on the PubMed website. The highest is about ten percent higher.

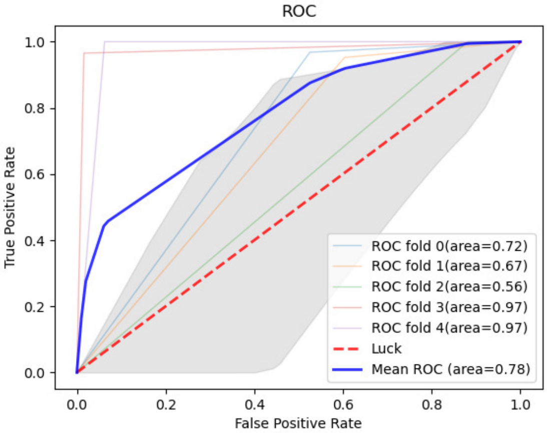

Figure 16. Svm-AdaBoost 5- fold cross validation roc curve.

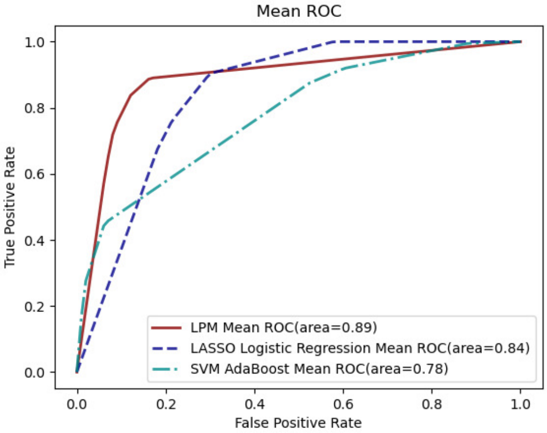

Figure 17. Mean ROC comparison.

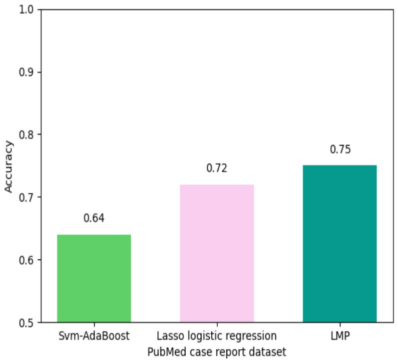

Figure 18. Comparison of accuracy in case report dataset.

As shown in Figure 19a, the accuracy on the electronic medical record dataset collected from a hospital is close, and the LMP algorithm is about one percent higher than the other two algorithms. As shown in Figure 19b, and the LMP algorithm is close to the LASSO logistic regression algorithm accuracy on the TCGA-CESC dataset, which is much higher than the SVM-AdaBoost algorithm. These can prove that the accuracy of our proposed LMP algorithm can achieve good results on different datasets.

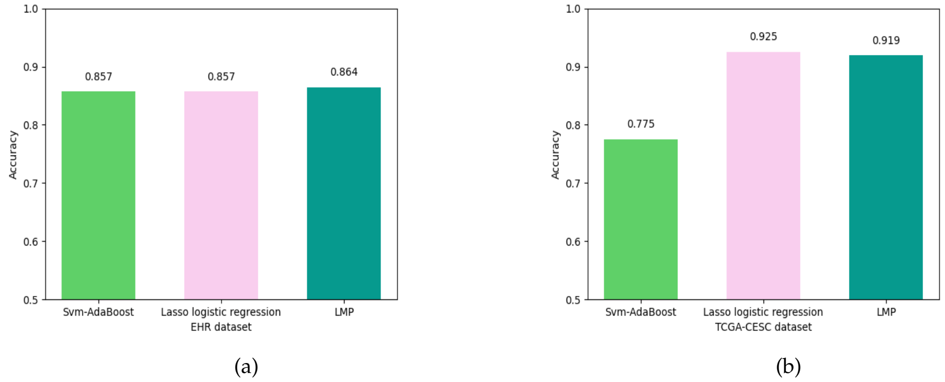

Figure 19. Accuracy Comparion with different datasets: EMR dataset (**a**) and TCGA-CESC dataset (**b**).

Before conducting the comparison experiments, we had to convert the original data in text form into key risk factors as attributes, and the corresponding values were taken as worthy of structured data, which took a lot of time and there were more missing values, which were filled as follows: categorical variables were filled with zeros, and continuous variables were filled with mean values. The LPM algorithm directly extracts the textual cases, which is easy to execute and time-saving, while the other method requires the conversion of the textual contents of the EMR data into tabular data form, which is tedious to organize manually. Secondly, the LPM algorithm can avoid the problem of many missing values in the case data, based on the network structure of risk factor association, and adopts the calculation idea of "what is available", using all the information mentioned in the case data to calculate and map to the set of key risk factors. The other algorithm requires all cases to have values relative to all attributes, and the treatment of missing values will directly affect the accuracy of the calculation.

To verify the validity of the new risk factors mined, we compared the prediction accuracy before and after the addition of the newly mined key risk factors, as shown in Figure 20.

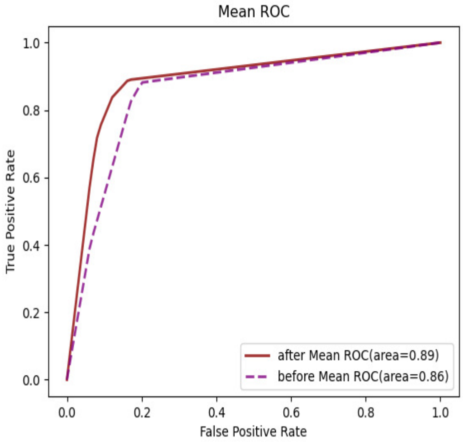

Figure 20. Comparison of LMP algorithm average roc before and after adding new digging factors.

The area under the curve increased after adding the newly mined factors, indicating the validity of our mined factors for disease prediction.

8. Discussion

New key risk factors in this study were confirmed to be associated with the development of cervical lesions or cervical cancer. The most common complaints of cervical cancer patients are excessive vaginal discharge, foul smelling, purulent or bloody vaginal discharge, and irregular vaginal bleeding, pruritus or abnormalities. The vagina is not a sterile environment. The vaginal microbiome is composed of many bacteria, including Lactobacillus and Gardnerella, etc. The vaginal pH < 7 is due to the role of Lactobacillus in the vagina, which can break down sugar and glycogen in the vaginal epithelium to produce acid and keep the vagina in an acidic environment, which is very important for maintaining the ecological balance in the vagina. Once the balance is disturbed, the pH of the vagina is altered and it becomes susceptible to vaginal HPV infection, leading to disease.

In addition, it has been found that estrogen, progesterone and human chorionic gonadotropin levels during pregnancy are positively correlated with human papillomavirus HPV 16 and HPV 18 infection, which indirectly suggests that pregnancy may promote the progression of cervical cancer. HPV infection also disrupts normal sex hormone function, diminishes the anti-estrogenic effect of progesterone in endometrial lesions, and increases the malignant transformation of cells.

It has also been shown that in general, women with good immune systems will heal themselves within a period of time after HPV infection, but persistent HPV infection will induce the onset of cervical lesions, and generally 1–2 years of persistent infection should be taken seriously.

The influence of emotional factors on the disease has also become increasingly significant in recent years, mainly because of the fast pace of contemporary life, high work pressure, prolonged anxiety or depression manifestations may cause a weak immune system, which in turn leads to HPV infection.

In the structure of the risk factor relationship diagram, for anxiety, depression, and immunity are all correlated with high-risk HPV infection, but in fact, the role of the relationship among them is that anxiety leads to depression, and these emotional factors of anxiety and depression affect the function of a person's immune system, which in turn leads to a decrease in immunity and leads to high-risk HPV infection.

As shown in Figure 21, there is a strong correlation between anxiety and depression, and there is also a relationship with immunity, and there is a strong correlation between immunity and high-risk HPV.

As shown in Figure 22, Lactobacillus, vaginal microenvironment, and immunity are all correlated with high-risk HPV infection in the structure of the risk factor relationship diagram, but in fact, the relationship between them is that Lactobacillus is one of the bacteria in the vaginal microenvironment, which plays a role in maintaining a normal vaginal microenvironment, but when it is abnormal, the balance of the vaginal microenvironment is disrupted and the ability to resist bacterial infection is reduced, and it also affects the immune system, which makes it easy to get high-risk HPV infection.

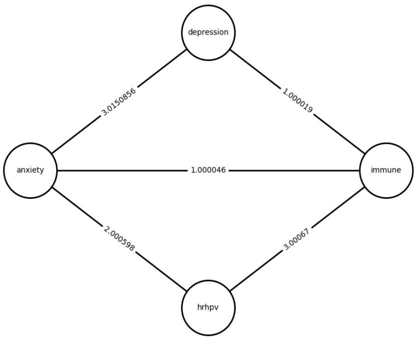

Figure 21. The relationship between anxiety, depression, immunity, and high-risk HPV.

As shown in Figure 23, in the structure of the risk factor relationship diagram, hormones, estrogen, progesterone, pregnancy, and immunity are all correlated with high-risk HPV infection, but in fact, the relationship between them is that estrogen is a type of hormone, and estrogen and progesterone are mutually regulated, and during pregnancy, there are dramatic and obvious changes in various hormones that can disrupt the immune system and weaken resistance, which can lead to high-risk HPV infection. This is the reason why many pregnant women are susceptible to HPV infection. The values of their relationship in our graph structure also indicate this close relationship. Jing Liu.et al Using statistical analysis method, we found that most of the cervical lesions during pregnancy were accompanied by high-risk HPV infection [37]. Our method, however, not only found that high-risk HPV infection and cervical lesions were associated with pregnancy, but also found the causative factors behind them, i.e., changes in hormone levels as well as changes in immunity.

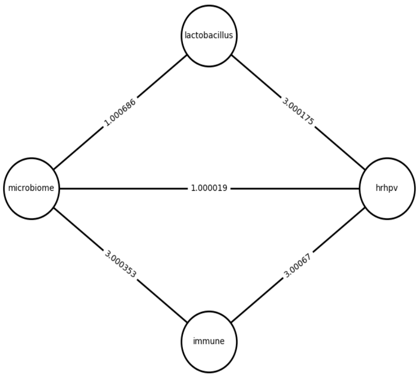

Figure 22. Relationship between Lactobacillus, vaginal microenvironment, immunity, and high-risk HPV types.

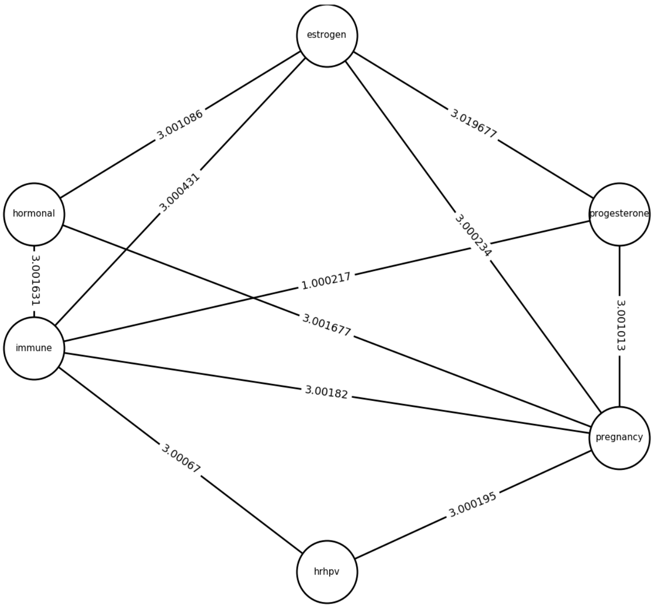

Figure 23. The relationship between pregnancy, hormones, estrogen, progesterone, immunity, and high-risk HPV.

9. Conclusions

In this study, the correlation of risk factors of cervical cancer was analyzed in depth, and the ontology knowledge base was established on this basis, which included all risk factors and their relationships. Then, this study proposed a key risk factors combination mining algorithm, which mined new key risk factors through common risk factors to obtain the new key risk factors of disease. These factors are important for the prevention of cervical cancer-related diseases. The experiments showed that the new key risk factors improved the accuracy of prediction of cervical lesion.

However, there is room for further improvement in our approach. When extracting risk factor keywords in the literature, we extracted all nouns and noun phrases in order not to miss them, which makes the graph size larger and leaves room for improvement in terms of runtime, which can be improved in the future by utilizing accurate natural language processing models to extract entities related to disease risk factors.

Author Contributions: Conceptualization, T.Z. and L.W.; formal Analysis, T.Z and P.X.; funding acquisition, L.W.; investigation, L.W. and P.X.; methodology, T.Z. and P.X.; software, P.X. and Y.T.; validation, P.X. and Y.T.; resources, L.W.; data curation, P.X.; writing-original draft, T.Z and P.X.; writing-review and editing, T.Z. and P.X.; visualization, P.X. and Y.T.; supervision, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No.62102076), and by the Science and Technology Development Plan of Jilin Province, China (No. 20220402033GH).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A publicly available dataset was analyzed in this study. The data carbe found here: (1) The Cancer Genome Atlas (TCGA) dataset: https://portal.gdc.cancer.gov/projects/TCGA-CESC (accessed on 31 August 2023). (2) The cervical cancer risk factor guidelines dataset: https://www.cancer.org/cancer/types/cervical-cancer/causes-risks-prevention/risk-factors.html (accessed on 3 January 2020). (3) The UCI case dataset: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors (accessed on 2 March 2017).

Acknowledgments: The authors would like to thank the reviewers for their essential suggestions for improving the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liverani, C.A.; Di Giuseppe, J.; Giannella, L.; Delli Carpini, G.; Ciavattini, A. Cervical cancer screening guidelines in the postvaccination era: Review of the literature. *J. Oncol.* **2020**, *2020*, 8887672 . [CrossRef]
2. Beardo, P.; Truan Cacho, D.; Izquierdo, L.; Alcover-Garcia, J.B.; Alcaraz, A.; Extramiana, J.; Mallofré, C. Cancer-specific survival stratification derived from tumor expression of tissue inhibitor of metalloproteinase-2 in non-metastatic renal cell carcinoma. *Pathol. Oncol. Res.* **2019**, *25*, 289–299. [CrossRef]
3. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
4. Yadav, S.K.; Akhter, Y. Statistical modeling for the prediction of infectious disease dissemination with special reference to COVID-19 spread. *Front. Public Health* **2021**, *9*, 645405. [CrossRef]
5. Pandey, R.S.; Srivastava, V.; Yadav, L.B. Research trends and solutions for secure traffic management of SDN. *Aptikom J. Comput. Sci. Inf. Technol.* **2017**, *2*, 97–105. [CrossRef]
6. Manur, M.; Pani, A.K.; Kumar, P. A prediction technique for heart disease based on long Short term memory recurrent neural network. *Int. J. Intell. Eng. Syst.* **2020**, *13*, 31–39. [CrossRef]
7. Adler, A. Using Machine Learning Techniques to Identify Key Risk Factors for Diabetes and Undiagnosed Diabetes. *arXiv* **2021**, arXiv:2105.09379.
8. Vijaya Saraswathi, R.; Gajavelly, K.; Kousar Nikath, A.; Vasavi, R.; Reddy Anumasula, R. Heart Disease Prediction Using Decision Tree and SVM. In *Proceedings of the Second International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 69–78.
9. Abdoh, S.F.; Rizka, M.A.; Maghraby, F.A. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access* **2018**, *6*, 59475–59485. [CrossRef]
10. Gupta, A.; Slater, J.J.; Boyne, D.; Mitsakakis, N.; Béliveau, A.; Druzdzel, M.J.; Brenner, D.R.; Hussain, S.; Arora, P. Probabilistic graphical modeling for estimating risk of coronary artery disease: Applications of a flexible machine-learning method. *Med. Decis. Mak.* **2019**, *39*, 1032–1044. [CrossRef]
11. Nie, B.; Li, C.; Wang, H. KA-NER: Knowledge Augmented Named Entity Recognition. In Proceedings of the Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, 4–7 November 2021; Proceedings 6; Springer: Berlin/Heidelberg, Germany, 2021; pp. 60–75.
12. Han, P.; Guo, J.; Lai, H.; Song, Q. Construction method of knowledge graph under machine learning. *Int. J. Grid Util. Comput.* **2022**, *13*, 11–20. [CrossRef]
13. Zhang, B.; Hu, Y.; Xu, D.; Li, M.; Li, M. SKG-Learning: A deep learning model for sentiment knowledge graph construction in social networks. *Neural Comput. Appl.* **2022**, *34*, 11015–11034. [CrossRef]

14. Ji, Z.; Shen, Y.; Sun, Y.; Yu, T.; Wang, X. C-CLUE: A benchmark of classical Chinese based on a crowdsourcing system for knowledge graph construction. In Proceedings of the Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, 4–7 November 2021; Proceedings 6; Springer: Berlin/Heidelberg, Germany, 2021; pp. 295–301.
15. Chang, D.; Chen, M.; Liu, C.; Liu, L.; Li, D.; Li, W.; Kong, F.; Liu, B.; Luo, X.; Qi, J.; et al. Diakg: An annotated diabetes dataset for medical knowledge graph construction. In Proceedings of the Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, 4–7 November 2021; Proceedings 6; Springer: Berlin/Heidelberg, Germany, 2021; pp. 308–314.
16. Wang, L.; Shan, M.; Zhou, T.H.; Ryu, K.H. Valuable Knowledge Mining: Deep Analysis of Heart Disease and Psychological Causes Based on Large-Scale Medical Data. *Appl. Sci.* **2023**, *13*, 11151. [CrossRef]
17. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
18. Alam, S.; Bakshi, B.; Maity, R.; Das, S.; Chaudhuri, A. Heart Disease Diagnosis and Prediction using Multi Linear Regression. *Int. J. Eng. Technol. Manag. Sci.* **2023**, *7*, 210–221.
19. Luo, J.; Yan, H.; Yuan, Y. Risk factors analysis and classification on heart disease. *Soft Comput.* **2020**, *24*, 13167–13178. [CrossRef]
20. Zhao, J.; Zhang, Y.; Qiu, J.; Zhang, X.; Wei, F.; Feng, J.; Chen, C.; Zhang, K.; Feng, S.; Li, W.D. An early prediction model for chronic kidney disease. *Sci. Rep.* **2022**, *12*, 2765. [CrossRef] [PubMed]
21. Christensen, T.; Frandsen, A.; Glazier, S.; Humpherys, J.; Kartchner, D. Machine learning methods for disease prediction with claims data. In Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 4–7 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 467–4674.
22. Swarupa, A.; Sree, V.H.; Nookambika, S.; Kishore, Y.K.S.; Teja, U.R. Disease prediction: Smart disease prediction system using random forest algorithm. In Proceedings of the 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, 13–14 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 48–51.
23. Yang, X.; Tong, Y.; Meng, X.; Zhao, S.; Xu, Z.; Li, Y.; Liu, G.; Tan, S. Online adaptive method for disease prediction based on big data of clinical laboratory test. In Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 26–28 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 889–892.
24. Farooqui, M.; Ahmad, D. Disease prediction system using support vector machine and multilinear regression. *Int. J. Innov. Res. Comput. Sci. Technol.* **2020**, *8*, 2347–5552. [CrossRef]
25. Faruque, M.F.; Asaduzzaman, A.; Hossain, S.M.M.; Furhad, M.H.; Sarker, I.H. Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation. *Eai Endorsed Trans. Pervasive Health Technol.* **2019**, *5*, e7. [CrossRef]
26. An, Y.; Huang, N.; Chen, X.; Wu, F.; Wang, J. High-Risk Prediction of Cardiovascular Diseases via Attention-Based Deep Neural Networks. *Ieee/Acm Trans. Comput. Biol. Bioinform.* **2019**, *18*, 1093–1105. [CrossRef]
27. Alaiad, A.; Hassan, N.; Mohsen, B.; Balhaf, K. Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease. *J. Inf. Knowl. Manag.* **2020**, *19*, 2040015. [CrossRef]
28. Luo, P.; Tian, L.P.; Chen, B.; Xiao, Q.; Wu, F.X. Ensemble disease gene prediction by clinical sample-based networks. *BMC Bioinform.* **2020**, *21*. [CrossRef] [PubMed]
29. Fan, L.; Hou, J.; Qin, G. Prediction of Disease Genes Based on Stage-Specific Gene Regulatory Networks in Breast Cancer. *Front. Genet.* **2021**, *12*, 717557. [CrossRef] [PubMed]
30. National Center for Biotechnology Information. Available online: https://pubmed.ncbi.nlm.nih.gov (accessed on 11 March 2024).
31. Genomic Data Commons Data Portal. Available online: https://portal.gdc.cancer.gov/ (accessed on 11 March 2024).
32. Risk Factors for Cervical Cancer. Available online: https://www.cancer.org/cancer/cervical-cancer (accessed on 11 March 2024).
33. Center for Machine Learning and Intelligent Systems. Available online: https://archive.ics.uci.edu/ml/datasets.php (accessed on 11 March 2024).
34. Yang, X.; Xiao, F. An improved gravity model to identify influential nodes in complex networks based on k-shell method. *Knowl.-Based Syst.* **2021**, *227*, 107198. [CrossRef]
35. Ahuja, Y.; Kim, N.; Liang, L.; Cai, T.; Dahal, K.; Seyok, T.; Lin, C.; Finan, S.; Liao, K.; Savovoa, G.; et al. Leveraging electronic health records data to predict multiple sclerosis disease activity. *Ann. Clin. Transl. Neurol.* **2021**, *8*, 800–810. [CrossRef]
36. Akbar, W.; Wu, W.p.; Saleem, S.; Farhan, M.; Saleem, M.A.; Javeed, A.; Ali, L.; Bashir, A.K. Development of Hepatitis Disease Detection System by Exploiting Sparsity in Linear Support Vector Machine to Improve Strength of AdaBoost Ensemble Model. *Mob. Inf. Syst.* **2020**, *2020*, 8870240. [CrossRef]
37. Liu, J.; Li, Y.; Bo, D.; Wang, J.; Wang, Y. High-risk human papillomavirus infection in pregnant women: A descriptive analysis of cohorts from two centers. *J. Investig. Med.* **2022**, *70*, 1494–1500. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.