IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 November 2023, accepted 24 December 2023, date of publication 26 December 2023,
date of current version 11 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347635

RESEARCH ARTICLE

Multidimensional Prediction Method for Thyroid Cancer Based on Spatiotemporally Imbalanced Distribution Data

ZHIWEI JIA[1], YUQI HUANG[1], YANHUI LIN[2], MIN FU[3], AND CHENHAO SUN[1]
[1]State Key Laboratory of Disaster Prevention and Reduction for Power Grid, Changsha University of Science and Technology, Changsha 410114, China
[2]Health Management Center, The Third Xiangya Hospital, Central South University, Changsha, Hunan 410013, China
[3]Department of Ophthalmology, Zhujiang Hospital, Southern Medical University, Guangzhou 510282, China

Corresponding author: Zhiwei Jia (jiayege@csust.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 52207074; and in part by the Research Foundation of Education Bureau of Hunan Province, China, under Grant 23A0255.

ABSTRACT In complex data environments, rational handling of unbalanced datasets is key to improving the reliability of early disease prediction. Early warning of disease risk in both temporal and spatial terms, contributes to disease prevention and treatment. To this end, a bi-dimensional substratum information mining model based on Association Rule Digging with Dynamic Thresholding and Weight Optimization (ARDdtwo) was proposed for the early diagnosis of thyroid cancer. It is an integrated assessment framework consisting of association rule digging by constructing a dynamic threshold model (ADRcdt) for qualitative analysis, and a self-optimizing component importance measurement model (SoCIM) for quantitative analysis. ARDcdt incorporates temporal and spatial features of sparse data to address the distributional bias problem. Moreover, new importance diagnostic calculations were designed to further identify high-risk low-frequency (HRLF). The SoCIM can determine the relative weight of each component by assessing its level of risk in the overall system based on the Risk Enhancement Level (REL) and Risk Reduction Level (RRL), realizing the self-adjustment and optimization of the weight setting. Finally, the model was validated through an empirical analysis. The evaluation of the research work shows that improved results were achieved, such as accuracy, f1-score, and precision, with optimized values of 36.04%,56.57%, and 53.89%, respectively. The overall area under the curve for the model was 0.882. This proves the validity of the proposed model for practical applications. For patients, it can simplify the pathological process and reduce the examination costs.

INDEX TERMS Disease early prediction, bi-dimensional substratum information mining, ARDdtwo, high-risk low-frequency.

I. INTRODUCTION

The thyroid gland is an important endocrine gland in the human body and is responsible for synthesizing, storing, and releasing thyroid hormones, which play a crucial role in metabolism, growth, and development, as well as temperature regulation [1]. Thyroid cancer is a malignant tumor that occurs in the thyroid tissue and has been increasingly diagnosed in recent years, posing a significant global public health challenge. According to the ''Cancer Facts and Figures'' report published in 2023 by the United States, thyroid cancer is the most common cancer of the endocrine system, accounting for approximately 92.4% of new cases of endocrine cancer in the country [2]. Although the mortality rate of thyroid cancer is low, as the disease progresses, it can invade surrounding tissues and organs and even metastasize to other parts of the body, such as the lungs and bones, causing more severe damage to the body [3].

The diagnosis of thyroid cancer typically requires a combination of various examination methods, such as ultrasound, fine-needle aspiration cytology (FNAC), blood tests, and radioactive isotope scanning. Ultrasound is currently the

The associate editor coordinating the review of this manuscript and approving it for publication was Giovanni Dimauro.

preferred method for diagnosing thyroid nodules because it avoids excessive puncturing of human cells. However, there have also been issues with the overdiagnosis and overtreatment of thyroid nodules. In actual diagnosis, thyroid nodules exhibit strong heterogeneity with uneven internal components, and benign nodules and malignant tumors can have overlapping ultrasound images. Additionally, the images may contain many artifacts and noise; therefore, the differences in experience and perception of different doctors can affect the accuracy and consistency of the judgement [4], [5], increasing the risk of invasive testing and treatment for patients. Similar to most diseases, people often lack concern about their thyroid health before symptoms appear. Furthermore, little research has been conducted on early prediction of thyroid disease. If the occurrence of the disease can be predicted before symptoms appear through existing medical data, it can prevent disease progression, take appropriate treatment measures earlier, and improve the cure rate.

Compared with traditional diagnosis based on doctors' knowledge and experience, machine learning utilizes a large amount of disease data for model training, thereby learning more complex and accurate disease prediction patterns and improving prediction accuracy. Commonly used machine learning methods include support vector machines (SVM), decision trees, and random forests. These methods can extract features and train models using existing disease data, and have been widely researched and applied in disease prediction, such as predicting the risk of diabetes, heart disease, and cancer [6], [7], [8].

In contrast to the aforementioned methods, association rules can discover hidden correlations between different variables that may not be easily observed directly or ignored. By mining these associations, useful information can be extracted and applied to prediction models to improve the prediction accuracy. Association rules can also efficiently process large volumes of data and extract meaningful association patterns from massive datasets. This helps expedite the construction and analysis of prediction models, thereby enhancing efficiency. Additionally, it is the result of data-based analyses that are objective and interpretable, which makes the results of the prediction model easier to understand and accept [9]. Therefore, this method is well suited for the research purpose of this study, which is to achieve large-scale early prediction of thyroid cancer from existing physical examination data that can be reasonably accessed.

The clinical diagnosis of thyroid cancer is complex and expensive. From the patient's perspective, improvements could be made in terms of reducing costs and easing the anxiety of waiting for results. As a result, we propose building an accurate and efficient early mass prediction system for thyroid cancer based on machine learning to help patients simplify the examination process and reduce the burden of visiting the doctor.

This paper facilitates our understanding of getting a real sense of the important role as well as the great potential of artificial intelligence in the field of medicine. The main contributions of this study are as follows.

(1) This model considers the unbalanced distribution of elements in time and space in a complex data environment, categorizes the elements into explicit and implicit elements by setting new thresholds on the input data in the time dimension, and further explores the implicit elements in the spatial dimension to determine potential high-risk low-frequency (HRLF) elements and obtain more accurate mining results.
(2) The Component Importance Measure (CIM) can be used to identify elements in the system that are at a higher risk of failure and take into account the relative weights of the implicit and explicit parts of each element, replacing the previous method of calculating the weights, which were based on the percentage of an element in the dataset.
(3) This model can find and visualize latent information that is not easily detectable from input data. It is adaptable in practical applications and can be used for early and large-scale prediction of different diseases, thereby reducing the cost of testing for patients.

The remainder of this paper is organized as follows. A review of related studies is presented in Section II. Section III provides a detailed representation of the model-building process, including data collection and processing, creation of ARDcdt and SOCIM, and selection of model evaluation criteria. Section IV presents an evaluation of the relevant results and discusses the implications of the study. Section V summarizes the study and proposes future work.

II. RELATED WORK

Both medical imaging and machine learning are widely used for the diagnosis of thyroid disorders. Given that conventional ultrasound characterization of thyroid cancer does not include either PPV or high sensitivity at the same time, Ho et al. used univariate analysis and multivariate logistic regression to construct a joint prediction model to identify specific risk factors for PTMC in female patients who have had children, which effectively improved diagnostic accuracy, but the scope of this study was limited, and the data sample was insufficient [16]. Li et al. used multiple linear regression analysis to develop a diagnostic model for differentiating thyroid cancer in clinical practice and developed the Thyroid Malignancy Scoring System (TMRS), given that the conventional diagnostic methods of ultrasonography and FNAB do not provide a definitive diagnosis of thyroid malignancy. However, the model is aimed at the diagnosis of thyroid cancer and cannot be used to screen for thyroid disease in the general population [17]. Nan Miles Xi et al. constructed a machine learning-based diagnostic model for thyroid cancer that combines machine learning with clinical data. A deep neural network model incorporating algorithms such as convolutional neural networks (CNN) and recurrent neural networks (RNN) was used to classify and predict thyroid cancer,

and experiments were conducted to demonstrate the model's potential for improving diagnostic accuracy. However, there were some limitations to this study's approach. For example, the method uses L1 regularization for feature selection, which may filter out certain features that are useful for diagnosis. Additionally, manual feature extraction is susceptible to subjective factors [18].

At present, few studies have specifically focused on the early prediction of thyroid cancer. The methods mentioned above are applicable for diagnosing the benign and malignant nature of thyroid nodules in patients but may not be suitable for large-scale early prediction in the general population. Association rule mining (ARM) is applicable to many fields of predictive modeling, especially in scenarios where it is necessary to mine the association patterns and discover hidden rules in data. It was originally designed by researchers to discover hidden relationships in market basket databases to improve efficiency [19]. It can mine valid pattern information from large amounts of data and is suitable for managing large datasets that rely on existing sample data for large-scale early prediction. Kang et al. proposed a satellite power system state prediction method based on online learning with parameter association rules, and an online rules-limited temporal convolutional network was proposed to improve the adaptability of the state prediction method. An association rule mining approach based on the fusion of Frequent Pattern Growth (FP-Growth) and trend sign aggregation approximation (TSAX) is used to monitor whether anomalies in the data may lead to unwanted model updates, and the results illustrate that the maximum mean absolute error of the proposed method comes from a shunt current of less than 0.42, which means that the effectiveness of the proposed method is ensured by the ability of association rules to effectively detect abnormal data [10]. Sheng et al. proposed a new approach for association rule mining in the context of power transformer state parameters using big data, which utilizes a probabilistic graph model to capture the relationships and dependencies among different state parameters of power transformers and developed a novel algorithm to mine association rules from a probabilistic graph model. The algorithm considers the probabilities and dependencies between parameters to identify meaningful and statistically significant associations, and the results demonstrate that the discovered association rules can provide valuable insights for condition monitoring, fault diagnosis, and maintenance decision-making in power transformer systems [11]. Yuan et al. proposed a method for predicting overall traffic modes using the VOMM (Vector Outer Product Model) approach and an AR (Association Rule) mining algorithm with large-scale data. This study utilized association rules to extract the traffic state relationship between different regions from historical data and discovered the correlation and patterns between traffic modes and various factors. The experimental results showed that the proposed method performed well in predicting overall traffic modes. The combination of VOMM and AR mining algorithms enables the model to capture the complex interactions among various factors and accurately predict the traffic mode [12]. Sun et al. proposed an ensemble system for predicting the spatial and temporal distribution of energy security weaknesses in transmission networks, established a fuzzy inference with a rare association rule learning system to predict the spatial and temporal distribution of energy security vulnerabilities in the long term, and demonstrated the practical application of this methodology to guarantee the sustainability and security of energy supply [22].

With the rapid development of information technology and application of big data, a large amount of medical data has been obtained. Making full use of these data resources, valuable information can be extracted to support medical decision-making, disease prediction, diagnosis, and treatment. Owing to the ability of association rules to handle large-scale data information as well as their interpretability, many studies have used it for disease prediction. For example, Khedr et al. presents an efficient method for association rule mining from distributed medical databases to predict heart diseases. A distributed association rule mining algorithm is used to perform data mining efficiently in distributed healthcare databases through parallel computing techniques. This algorithm reduces computational complexity and improves mining efficiency [13]. Tandan used association rules to mine symptom patterns and overall symptom rules in patients with COVID-19 to explore common symptoms and differences in symptoms between different types of patients [20]. Yujie et al. constructed an early prediction and diagnosis model for coronary heart disease based on the WEKA data mining platform, first using the J48 decision tree to construct a coronary heart disease diagnosis prediction model, then mined strong association rules through the Apriori algorithm, and finally quantified the influence of input attributes on coronary heart disease diagnosis according to the Multilayer Perceptron algorithm; however, the model was established with too few examples, and the accuracy of the model was not good enough [21]. Li et al. applied data mining techniques to the risk prediction of diabetes mellitus and conducted experiments on medical datasets. The relationship between diabetes and various symptoms was found to provide effective support for early risk prediction of diabetes [14].

Although ARM has been widely used for predictive modeling, it has some limitations. In practical applications, for the elements that appear in biased data with low frequency, the traditional algorithm uses a fixed scoring method to analyze all the variables, which makes the model pay too much attention to the general elements that occupy the main position in the dataset and makes these special elements directly discarded without being analyzed, but these elements may contain high-risk elements [22]. When quantifying the importance of elements, existing algorithms usually determine their weights solely based on the proportion of elements in the system as a measure, and the parameters are not adjusted and optimized during the model-building process; thus, the

importance of sparse data is not objectively analyzed. Based on these problems, this paper proposes a bi-dimensional substratum information mining model, namely, Association Rule Digging with Dynamic Thresholding and Weight Optimization Model (ARDdtwo).

First, association rule digging with the construction of a dynamic threshold model(ARDcdt) is designed to incorporate sparse data into the assessment, and the conditional importance thresholds and diagnostic criteria score calculation methods are improved to analyze the potential implicit relationships in complex data environments using a temporal-spatial-mining system, which not only focuses on the general elements that dominate the dataset but also fully analyzes and evaluates the sparsely distributed elements. Considering the degree of influence of age on thyroid lesions, the calculation method of the threshold value of the diagnostic criteria was reset with the age distribution as the division criterion, and specific threshold values were used for information mining for different age groups. In the spatial dimension, a new calculation method of the conditional importance diagnostic criteria was designed, which can further identify the elements with high risk–frequency (HRLF)for the entire system from the set of implicit elements. Finally, to adjust and optimize the parameter weights, the weight measurement method self-optimizing component importance measurement model (SoCIM) is designed based on the component importance measure (CIM), which can realize the evaluation criteria of weights based on the assessment of the risk index of each element.

III. MATERIALS AND METHODS

A. DATA PRE-PROCESSING

Because a large amount of real data is essential to build a reliable prediction model, this study collected medical examination data for 2021 from 11 provinces provided by The Third Xiangya Hospital. Considering the large number of medical examination forms and the redundancy of information in the forms, we pre-processed the obtained data. Outliers due to manual entry or improper operation of the system, as well as some samples with a large number of missing and duplicate values, were removed, and the continuous data in the forms were discretized to facilitate final result analysis and model building. A total of 1456 disease samples were obtained from 17817 medical examination data, and the elements under each feature in the resulting data samples were analyzed. The age range of the sick samples was 0-90 years old. To target the causative factors of different age groups, we divided the sick samples into three groups according to their age, and the distribution is shown in Fig.1.

A standardized database allows all the data to be grouped in a uniform space to perform the required processing steps. Therefore, in this study, let $P_t \in P = \{P_1, P_2, \ldots P_z\}$ represent the medical examination data entered into database P for a certain age group. In P_t, let $U = \{u_1, u_2, \ldots, u_n, \ldots, u_Y\}$ be a set containing all the features contained in the database,

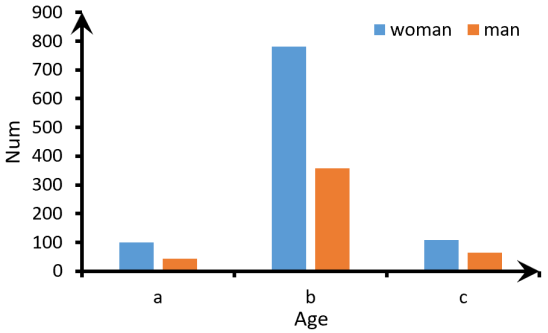


FIGURE 1. Distribution of the total number of male patients and the total number of female patients with thyroid cancer in each age group from the valid data screened in 11 provinces in 2021. (a) 0-30 years. (b) 31-60 years. (c) 61-90 years.

and u_n be one of the relevant feature variables, Y be the corresponding target feature variable (i.e., the diagnostic result for that piece of data). Then, $Y = \{Y_1, Y_2, \ldots, Y_i\}$ is a set containing all the target variables. Each feature u_n in the database consists of a set of elements, $a_{n,1}, a_{n,2}, \ldots, a_{n,k}$, which are also known as items in the association rules. In this study, an element is a component of a characteristic (e.g., 'does not smoke but often smokes secondhand' is an element of the characteristic 'does smoke'). Let $I = \{b_1, b_2, \ldots, b_n\}$ be a set containing all input variables; then, $a_{n,k}$ is equivalent to any one of the variables in I. Thus, the set of items X is a subset in I. The association rule can be expressed as $X \rightarrow Y$. To facilitate the processing of data during modeling by combining the medical examination data numbers with the data processing space P, it can be denoted as P_l. Based on the above settings, each set is written as a matrix, and the data processing space used to mine implicit variables can be represented as

$$P_l = \begin{bmatrix} l_1 & \\ l_2 & P \\ \vdots & \\ l_i & \end{bmatrix} = \begin{bmatrix} & u_1 & u_2 & \cdots & u_n & \cdots & u_Y \\ l_1 & b_{11} & b_{12} & \cdots & b_{1n} & \cdots & Y_1 \\ l_2 & b_{21} & b_{22} & \cdots & b_{2n} & \cdots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ l_i & b_{i1} & b_{in} & \cdots & b_{in} & \cdots & Y_i \end{bmatrix} \quad (1)$$

Starting from the second row, each row represents the corresponding physical examination data of the examiner and b_{ij} represents the elements contained under feature u_j, l_i represents the ith row of medical examination data in the data sample.

B. FEATURES SELECTION

Because the complete medical examination form contains 43 features, the data dimension is too high for modeling. Therefore, the original features must be filtered to improve the efficiency and flexibility of the model. First, we removed features that were not relevant to model building such as mobile phone numbers, dates of medical examinations, and payment methods. After filtering out this invalid information, we used XGboost to perform feature importance ranking

to achieve dimensionality reduction for subsequent model building [23], [24]. The designed filtering process is illustrated in Fig.2.

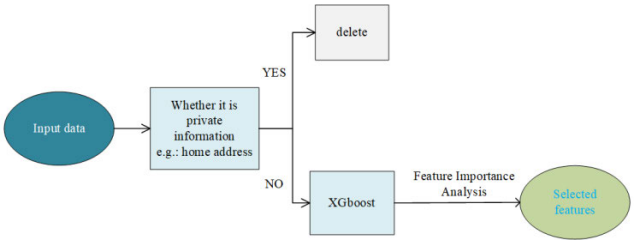


FIGURE 2. Flow chart of input features selection.

C. ASSOCIATION RULE DIGGING WITH CONSTRUCTING DYNAMIC THRESHOLDS MODEL

1) CONDITIONAL IMPORTANCE DIAGNOSTIC THRESHOLD SETTING BY TIME DOMAIN

Traditional association rule mining uses a fixed importance evaluation criterion score calculation method to obtain potential relationships between elements and results in one dimension. However, for medical examination data and most medical data, the episodic nature of the disease determines the imbalance of the data; therefore, using a traditional algorithm will make the mining results biased toward the side with more quantity, which is unfair, and a low percentage of easily overlooked factors is likely the cause of the lesions. Moreover, there are differences in the probability of the same factor causing lesions in different populations; therefore, the data must be thoroughly analyzed. For example, Fig.2. shows that there is a clear difference in the number of diseases between young and middle-aged populations, which means that the probability of disease differs between these two groups. Using the same threshold for mining analysis will lead to the youth group's pathogenic elements not being paid attention to because they are lower than the set threshold, and the screening results will be biased toward the middle-aged group's pathogenic factors, which will reduce the accuracy of the group's prediction. However, the prevalence rate is increasing in the youth group and their pathogenic factors should be analyzed as well. Therefore, we need to improve the threshold setting method so that it is more reasonable to analyze sparsely distributed data comprehensively.

Considering the relationship between age and the probability of disease occurrence, we categorized the data by age group and designed a method for setting the threshold for the diagnostic criteria of conditional importance. It enabled us to set the corresponding thresholds according to the distribution of patients in each age group in a targeted manner, thus, people in the age group with the lowest probability of disease could also be adequately analyzed. Our data sample was divided into three parts: a represents young population (18-30 years old), b represents middle-aged population (31-60 years old), and c represents elderly population (61-90 years old). One age group was chosen as the base unit period, and the same threshold was used for data samples within the same age group.

According to the distribution of patients in each age group, this study proposes five corresponding time-domain-based importance diagnostic calculations, which can be expressed as follows: (2)–(6), as shown at the bottom of the next page, where $\langle \ldots \rangle$ represents the number of diseased samples in the database that satisfy all the conditions included in the equation simultaneously, the subscript '0' represents the initial preset threshold, $Y(h_y)$ represents the age group in which the diseased sample is located, and $Y(h_y^{\max})$ represents the age group with the highest number of diseased samples in the database.

2) SIGNIFICANCE DIAGNOSIS METHOD DESIGN BY SPATIAL DOMAIN

a: TRADITIONAL DIAGNOSTIC CRITERIA

The expression for traditional association rules can be expressed as $X \rightarrow Y$ (X and Y are the precursors and successors of the association rules). It has five important diagnostic criteria: support, confidence, conviction, lift, and leverage. Assuming that the total amount of data is D, they can be expressed as

Support ($sup(X \rightarrow Y)$) is the probability of occurrence of item sets X and Y in the total item set; it is usually used to remove meaningless rules.

$$\sup(X \rightarrow Y) = \frac{count(X \cup Y)}{D} \tag{7}$$

Confidence ($conf(X \rightarrow Y)$) is the ratio of the probability of occurrence of X and Y simultaneously to the probability of occurrence of X alone, which reflects the reliability of the rule.

$$conf(X \rightarrow Y) = \frac{\sup(X \cap Y)}{\sup(X)} \tag{8}$$

Lift ($lift(X \rightarrow Y)$) is complementary to confidence and denotes the enhancement of the probability of occurrence of Y by the occurrence of X.

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{\sup(Y)} \tag{9}$$

Conviction($conv(X \rightarrow Y)$) analyses the situation when there is no occurrence of Y. It represents the product of the two probabilities of X appearing and Y not appearing, and the ratio between the probabilities of X appearing with Y not appearing.

$$conv(X \rightarrow Y) = \frac{1 - \sup(Y)}{1 - conf(X \rightarrow Y)} \tag{10}$$

Leverage($leve(X \rightarrow Y)$) denotes the probability of X and Y appearing in D simultaneously when X and Y are not completely independent.

$$leve(X \rightarrow Y) = \sup(X \rightarrow Y) - \sup(X)\sup(Y) \tag{11}$$

b: CONDITIONAL IMPORTANCE DIAGNOSTIC CRITERIA

Data mining using the new thresholds set in the time domain yields implicit and explicit factors corresponding to each age group. Considering the rigour of disease diagnosis, it was possible that rare elements, which make up a small percentage of the database, may also be important to the occurrence of the disease.It is necessary to perform latent relationship mining among the implicit elements screened to identify the high-risk-low-frequency (HRLF) elements.

Existing models suffer from the problem of considering dominant elements in the database but filtering out HRLF variables directly when analyzing potential relationships between variables. This is because the ODM still uses the same fixed thresholds to calculate the importance scores of the implicit variables in different features as the common variables in the corresponding features, resulting in a set of implicit variables being excluded directly. They did not reach the initial threshold set and their importance was not confirmed. For example, in one of the patient data in the database, ''positive optimism'' in the feature ''daily psychological state'' is the explicit factor, and ''negative depression'' is the implicit factor. When using traditional importance diagnostics, because its results are influenced by a large set of dominant variables in the database, sample data containing the implicit element ''negative low'' will be directly excluded because they do not reach a set fixed threshold.

This study resets the calculation method of important diagnosis based on the spatial domain to identify HRLF variables that lead to disease occurrence from the already screened hidden variables, and the association rule can be expanded as

$$X^A + X^B \to Y \tag{12}$$

where X^A represents the set of explicit variables and X^B represents the set of implicit variables.

A second correlation analysis of the database can be completed according to the distribution of the hidden variables in the distinctive features. When the feature contains a hidden element, the threshold setting method for the corresponding evaluation criterion can be written as (13)–(17), shown at the bottom of the next page, where x represents a value between 2 and i, y represents a value between 2 and n, T_N represents a range of values from 2 to $(n+1)$, $Y(h_t)$ represents one of the age bands a, b, and c.

D. SELF-OPTIMIZING COMPONENT IMPORTANCE MEASUREMENT MODEL

1) COMPONENT IMPORTANT MEASURE

In previous models, a common method for measuring the importance of elements was based on their proportion in the overall data or frequency of occurrence. However, this is not reasonable for data such as healthcare data, where there is often an extreme imbalance in outcomes, as the proportion of elements or frequency of occurrence does not equate to the degree of significance, so we need to devise a more scientific and rational method of calculating relative significance.

In this regard, the CIM principle can be used to measure the impact of different elements on the outcome, enabling the assessment of each element of the system based on its respective risk of failure, uncovering components that are prone to system failure, and improving system stability. The CIM computational model is usually a combination of the Risk Enhancement Level (REL) and the Risk Reduction Level (RRL) to assess the level of risk. REL is defined as the relative increase in the risk of causing disease when an element $a_{n,k}$ is present and is used to identify the components that maintain the stability of the system; RRL is defined as the

$$\min \sup = \min \sup_{0} \cdot \frac{\left\langle l_i \in P_l(x,1); P_l(x,n+2) = Y(h_y) \right\rangle}{\left\langle l_i \in P_l(x,1); P_l(x,n+2) = Y(h_y^{\max}) \right\rangle}, \quad x \in (1,i) \tag{2}$$

$$\min conf = \min conf_0 \tag{3}$$

$$\min conv = \min conv_0 \cdot \frac{\left\langle l_i \in P_l(x,1); P_l(x,n{+}2) = Y(h_y); P_l(x,n+1) = Y(h_i) \right\rangle}{\left\langle l_i \in P_l(x,1); P_l(x,n{+}2) = Y(h_y^{\max}) \right\rangle} \tag{4}$$

$$\min lift = \min lift_0 \cdot \frac{\left\langle l_i \in P_l(x,1); P_l(x,n+2) = Y(h_y); P_l(x,n+1) = Y(h_i) \right\rangle}{\left\langle l_i \in P_l(x,1); P_l(x,n+2) = Y(h_y^{\max}); P_l(x,n+1) = Y(h_i) \right\rangle} \tag{5}$$

$$\min leve = \min leve_0 \cdot \frac{\left\langle l_i \in P_l(x,1); P_l(x,n+2) = Y(h_y); P_l(x,n+1) = Y(h_i) \right\rangle}{\left\langle l_i \in P_l(x,1); P_l(x,n+2) = Y(h_y^{\max}); P_l(x,n+1) = Y(h_i) \right\rangle} \tag{6}$$

relative degree of reduction in the risk of pathogenicity when an element is not present and is used to identify the elements that reduce the occurrence of risk. The two are functionally complementary and can therefore be used to measure the trend and influence of the various components of the system on the occurrence of the disease. Their corresponding mathematical expressions are:

$$T^{REL}(a_{n,k}\,|r_i) = \frac{1 - R(0_K, Q(t_i))}{1 - R(Q(t_i))} \tag{18}$$

$$T^{RRL}(a_{n,k}\,|r_i) = \frac{1 - R(Q(t_i))}{1 - R(1_K, Q(t_i))} \tag{19}$$

where $1 - R(0_K, Q(t_i))$ represents the risk of disease when an element is determined to be present, $1 - R(1_K, Q(t_i))$ represents the relative degree of reduction in the risk of disease when an element is not present, and $1 - R(Q(t_i))$ represents the risk of the disease appearing in the system.

2) RELATIVE WEIGHT

Based on the CIM, it can analyze the impact of different elements on the risk of disease and thus construct a model for calculating the Elemental Spatial Risk Index (ESRI). The relative weights of the elements are measured by the degree of direct impact on the overall risk of the system at the time of their occurrence, which provides a more precise result of the weights of each element. The ESRI was calculated based on the REL and RRL. The ESRI of an element $a_{n,k} \in u_n$ in the system is denoted as $\omega_{a_{n,k}}$. From the above analysis, each element has a dominant and recessive part because the same element affects different age groups differently, and the relative weight for any element contains both the impact weight of the common set of variables and the impact weight of the HRLF. u_n^A is the subset that contains the explicit elements of all features and u_n^B is the set that contains the implicit elements. The total number of cases is represented by V. It can be expressed using the following formula:

$$\omega_{a_{n,k}} = \omega_{n,k}^A + \omega_{n.k}^B \tag{20}$$

where $\omega_{n,k}^A$ represents the risk created by explicit elements and $\omega_{n.k}^B$ represents the risk caused by implicit elements.

$$\omega_{n,k}^A = \begin{cases} 0, & a_{n,k} \in u_n^B \\ \sum_{x=2}^{|l_i \in P_l|} \frac{\langle P_l(x,y) = a_{n,k} \rangle}{\langle V \rangle}, & a_{n,k} \in u_n^A \end{cases} \tag{21}$$

According to risk structure theory, the overall risk of a system is related to the relative position of the elements in the system concerning their constituent structures. All the constituent parts in the database are independent of each other. The overall pathogenic risk of the system can be expressed as

$$O_S = \prod_{i=1}^{n} O_i \tag{22}$$

where O_S represents the risk of the system as a whole and O_i represents the risk of a component of the system. To solve for the overall risk of failure of the system, the risk structure of the system must be determined; therefore, the logical relationships between the features must be analyzed. For disease samples in the input data, even the absence of the corresponding element under any of the features may result in the disease no longer occur; therefore, in this system, the features should be connected in series. Assuming that each element is relatively independent, the overall risk of the system is obtained by the product of the risk of the corresponding elements in all features. Thus, the risk index from the implicit element can be expressed as

$$\omega_{n,k}^B = \begin{cases} 0, & a_{n,k} \in u_n^A \\ \omega_1 \cdot C_{REL} + \omega_2 \cdot C_{RRL}, & a_{n,k} \in u_n^B \end{cases} \tag{23}$$

$$Sup(X^A + X^B \to Y) = \frac{\langle l_i \in P_l(x,1); X^A \subseteq P_l(x,T_N) \neq \emptyset; P_l(x,y) \in X^B \neq \emptyset \rangle}{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset \rangle} \tag{13}$$

$$Conf(X^A + X^B \to Y) = \frac{\langle l_i \in P_l(x,1); X^A \subseteq P_l(x,T_N) \neq \emptyset; P_l(x,y) \in X^B \neq \emptyset; P_l(x,n+1) = Y(h_t) \rangle}{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset \rangle} \tag{14}$$

$$Conv(X^A + X^B \to Y) = \frac{1}{1 - Conf} \cdot \{1 - \frac{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset; P_l(x,n+1) = Y(h_t) \rangle}{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset \rangle}\} \tag{15}$$

$$lift(X^A + X^B \to Y) = Conf \cdot \frac{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset \rangle}{\langle l_i \in P_l(x,1); P_l(x,y) \neq \emptyset; P_l(x,n+1) = Y(h_t) \rangle} \tag{16}$$

$$leve(X^A + X^B \to Y) = \frac{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset; X^A \in P_l(x,T_N); P_l(x,n+1) = Y(h_t) \rangle}{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset \rangle} - Sup \cdot \frac{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset; P_l(x,n+1) = Y(h_t) \rangle}{\langle l_i \in P_l(x,1); P_l(x,y) \in X^B \neq \emptyset \rangle} \tag{17}$$

$$C_{REL} = \frac{1 - \prod_{K=1}^{L} \left\langle \sum_{x=2}^{\left|l_x \in P_y^B\right|} \frac{\left|l_i \in P_l(x,1); P_l(x,y) \in a_{n,k}; P_l(x,y) \in u_y^B\right|}{\left|l_i \in P_l(x,1); P_l(x,y) \in u_y\right|} \right\rangle}{1 - \prod_{y=2}^{n+1} \left\langle \sum_{x=2}^{\left|l_x \in P_y^B\right|} \frac{\left|l_i \in P_l(x,1); P_l(x,y) \in a_{n,k}; P_l(x,y) \in u_y^B\right|}{\left|l_i \in P_l(x,1); P_l(x,y) \in u_y\right|} \right\rangle} \tag{24}$$

$$C_{RRL} = \frac{1 - \prod_{K=1}^{L} \left\langle \sum_{x=2}^{\left|l_x \in P_y^B\right|} \frac{\left|l_i \in P_l(x,1); P_l(x,y) = a_{n,k}; P_l(x,y) \in u_y^B\right|}{\left|l_i \in P_l(x,1); P_l(x,y) \in u_y\right|} \right\rangle}{1 - \prod_{y=2}^{n+1} \left\langle \sum_{x=2}^{\left|l_x \in P_y^B\right|} \frac{\left|l_i \in P_l(x,1); P_l(x,y) \neq a_{n,k}; P_l(x,y) \in u_y^B\right|}{\left|l_i \in P_l(x,1); P_l(x,y) \in u_y\right|} \right\rangle} \tag{25}$$

where P_y^B represents a sub-matrix consisting of implicit elements, u_y^B stands for a subset of implicit elements by.

E. MODEL EVALUATION

The confusion matrix is commonly used as a metric to assess the performance of the model and is analyzed in matrix form by judging the sample data in the dataset according to the true categories and categories predicted by the classification model. Model evaluation metrics included accuracy, precision, sensitivity, specificity, and F1-score.

When assessing differences in model performance, the subject operating characteristic curve (ROC curve) is the core metric, and the classification model can be selected based on the performance of FP and TP metrics, as shown in Fig 3, Diagram A. The performances of different models can be visually compared using the area under the curve (AUC). When the curve is closer to the upper-left corner, the model's performance improves. However, the performance of a model cannot be fully measured using only precision and recall rates corresponding to a point. More comprehensive assessment of the model can be made employing a Precision-Recall (P-R) curve, as shown diagram B, which uses 'Recall' as the horizontal axis and 'Precision' as the vertical axis. This shows the balance between the precision and recall rates at different threshold values.

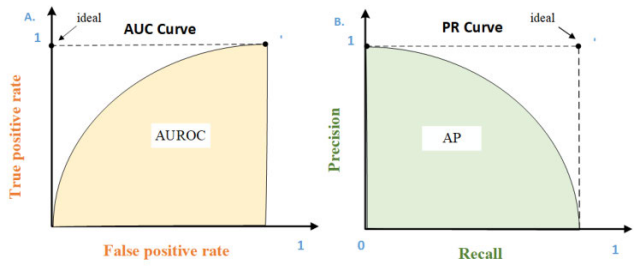


FIGURE 3. The ROC and PR curves. Diagram A represents the ROC curve and Diagram B represents the PR curve.

Furthermore, 5-fold cross-validation methodology was adopted in this study. Following five rounds of testing, four of the five folds were used as training data, and the remaining fold was used as validation data, thus ensuring that every record entered the database was validated at least once, which can further reduce the impact caused by potential data bias.

IV. REAULTS AND DISCUSSION

A. RESULTS OF FEATURE SELECTION

Considering that Body Mass Index (BMI) is more scientific than height and weight as separate assessment characteristics, 16 discrete and continuous features were obtained by the feature importance results derived from Xgboost, as shown in Table.1.

B. RESULTS OF THE MEANINGFUL IMPLICIT ELEMENT SELECTION

The initial threshold values were set to $minsup = 0.2$, $minconf = 0.6$, and $minconv = 1.1$. With the newly defined thresholds and conditional diagnostic criteria, the association rule mining results for each age group obtained after analyzing the input data are shown in Table.2. Notably, the thresholds and HRLF change with age.

Each group has its own HRLF elements, which suggests that many of the factors discarded when mining with ODM do not have a significant association with disease in that age group. When mining for deep hidden information. It can identify elements in the data that may have a low occurrence rate, but possess a strong influence on the outcome or target variable. These elements may be overlooked by traditional statistical methods that focus on more prevalent factors. Then, using (20), we calculate the weights that each element holds in the overall system. By utilizing ARDdtwo, the algorithm can assign different weights to each feature based on their importance in predicting the target variable.Through assigning higher weights to these influential but rare elements, it may lead to the discovery of previously unrecognised patterns, relationships or anomalies in the data.

The youth group had better physical health and emotional regulation and therefore had the lowest number of HRLF. The results show that drinking too much and sleeping too much are potentially high-risk factors. In the middle-aged population, it can be seen that there are more HRLF factors than in any other group, this is why people in this age group have the highest probability of developing the disease. Unlike the other two age groups, family history has a potentially high-risk impact on the ability to develop thyroid disease, especially hypertension. In addition, stress and psychological conditions in daily life have a potential impact on thyroid gland health. This may be related to the nature of work and stressful life of the middle-aged population, and the accumulation of unhealthy habits over time is responsible for the increased incidence of thyroid cancer. To reduce the incidence of thyroid cancer in the middle-aged population, it is important to reduce the intake of alcohol and tobacco in daily life, eat breakfast on time, sleep neither too long nor too short, preferably between 6-8 hours, exercise, relax, reduce

TABLE 1. Feature screening results.

Discrete features	Elements contained
Age	0-30 years(a); 31-60 years(b); 61-90 years(c)
Gender	Man(a); Woman(b)
Family history-parents	Diabetes(A); High blood pressure(B); Lipid abnormalities(C); heart disease(D); Chronic obstructive pulmonary disease(E); Hepatitis(F); Cerebrovascular diseases(G); Mental illness(H); Malignant tumors(I); No family history(N)
Family history-brothers and sisters	Diabetes(A1); High blood pressure(B1); Lipid abnormalities(C1); heart disease(D1); Chronic obstructive pulmonary disease(E1); Hepatitis(F1); Cerebrovascular diseases(G1); Mental illness(H1); Malignant tumors(I1); No family history(N)
Smoking status	no smoking(N); Non-smoker but often smokes second-hand smoke(a); smoking(b); Quit smoking(c)
Monthly alcohol consumption	Mild alcohol consumption(a); Moderate drinking(b); Heavy drinking(c); no alcohol(N)
Eating habits	Balanced meat and vegetables(a); Vegetarian-based(b); Meat-based food(c)
Taste Preference	Normal(a); heavy(b)
To have breakfast or not	Yes; No
Work/life stress	Little pressure(a); Less pressure(b); Moderate pressure(c); A lot of pressure(d);Higher pressure(e)
mental	Positive and optimistic(a); Negative depression(b); Calm and not susceptible to external changes(c); Vulnerable to change by external influences(d)
Continuous features	**Elements contained**
Naptime	0-0.5h(a);≥0.5h(b);No napping(N)
Length of sleep at night	Less than 6h(a);6-8h(b); More than 8h(c)
Number of exercises per week	No exercise(no);1-3times(a); more than 3 times(b)
Duration of a single exercise session	0-30min(a);30-60min(b);>1h(c); no exercise(N)
BMI	18.5-23.9(a); >23.9(b); <18.5(c)

Content () represents the representative name of the element in the database.

TABLE 2. Data mining results of ARDdtwo by each age group.

	Age -a		Age-b		Age-c	
Features	Explicit elements	HRLF	Explicit elements	HRLF	Explicit elements	HRLF
Gender	a; b	none	a; b	none	a; b	none
Family history-parents	A; B; C; N	none	N	B	A; B; I; N	none
Family history-siblings	N	none	N	A1; B1; C1; D1; E1; F1; G1	B1; C1; N	none
Smoking status	a; b; N	none	N; a	b	a; c; N	b
Monthly alcohol consumption	a; N	b; c	N	a; b	a; N	b; c
Eating habits	a; b; c	none	a	b; c	a; b; c	none
Taste Preference	a; b	none	a; b	none	a; b	none
To have breakfast or not	Yes; No	none	Yes	No	Yes; No	none
Length of sleep at night	a; b	c	b	a; c	a; b	c
Naptime	a; b; N	none	a; b; N	none	a; b; N	none
BMI	a; b; c	none	a; b	c	a; b	c
Number of exercises per week	a; b; N	none	a; b; N	none	a; b; N	none
Duration of a single exercise session	a; b; N	none	a; b	N	a; b; c; N	none
Work/life stress	a; b; c; d; e	none	a; c	b; d; e	a; b; c; e	d
mental	a; b; c; d	none	a	c; d	a; c; d	b

work and life stress, and develop good lifestyle habits. For the elderly population, smoking, drinking, stress, and depression were hidden factors mined by ARDdtwo.

Therefore, we need to care more about older adults to enable them to maintain a happy and positive mood in their daily lives. For themselves, they need to improve their living habits, not to smoke or drink, to maintain a balanced diet and exercise to keep their BMI within the normal range, in addition to not sleeping for too long at night.

C. MODEL COMPARISON

1) COMPARISON WITH TRADITIONAL MODELS

To further evaluate the performance of our system, we compared the designed model with a traditional association-rule model using the same selection of features. As shown in Fig.4, the AUC obtained with ODM was 0.541, whereas that obtained with ARDdtwo was 0.882 (a 63.03% increase). Moreover, ARDdtwo is not only closer to the upper left corner of the ROC graph but also has a higher AUC score,

which intuitively shows that the model designed in this study performs significantly better.

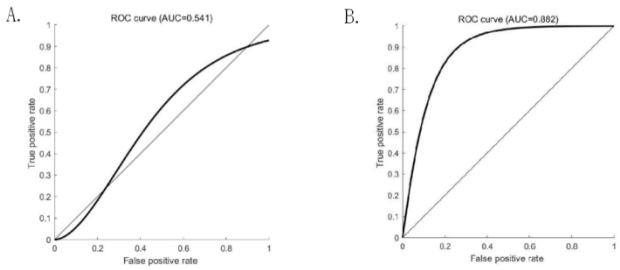


FIGURE 4. Results of AUC score. A represents ODM and B represents ARDdtwo.

Finally, in Fig.5, a line graph of the evaluation metric scores shows that ARDdtwo outperforms ARDdtwo in every aspect. The AUC, SE, accuracy, precision, recall, and F1-score were improved by 63.03%, 36.31%, 36.04%, 53.89%, 63.74%, and 56.57%, respectively. Therefore, the model designed in this study exhibited a significant effect enhancement.

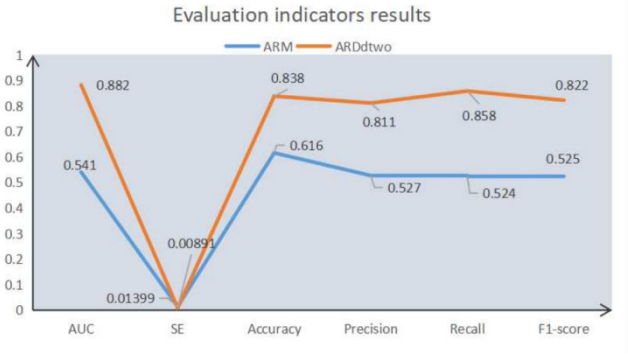


FIGURE 5. Evaluation results of each metric for ODM and ARDdtwo.

The number of negative samples was much larger than the number of positive samples in the medical examination sample data, we used P-R curves to compare the two models. The results are shown in Fig. 6, where Model_1 represents the ODM and Model_2 represents ARDdtwo. The area enclosed by the PR curve and the coordinate axis of ODM is 0.311, whereas the area enclosed by ARDdtwo is 0.723. In addition, Model_2 completely encloses Model_1, which shows that ARDdtwo performs much better for positive examples in the case of a data imbalance.

2) COMPARED TO THE EXISTING MODEL

The predictive model proposed in the literature [25] for patients with TI-RADS grade 4a thyroid nodules predicts the nature of their thyroid nodules. The selected characteristics were age and PLR, TSH, and ALB levels. The literature [26] used several machine-learning methods to predict thyroid cancer, and the results showed that Random Forest (RF) has the best overall prediction with an accuracy of up to 0.9091.

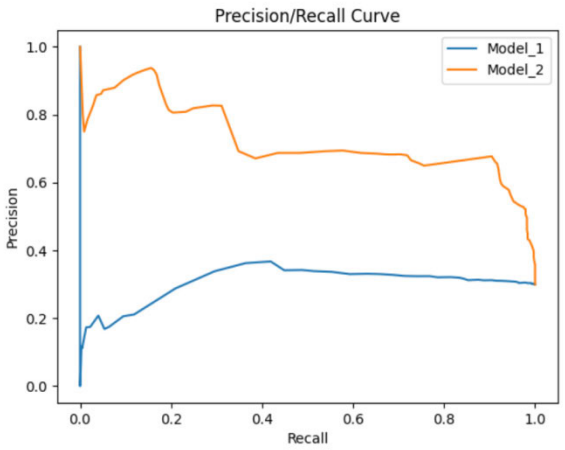


FIGURE 6. Graphs of P-R curves obtained from ODM and ARDdtwo for the same input data.

TABLE 3. Input feature for different models.

Author	Features
Rui Wang [25]	age TSH ALB age、TSH、ALB
Sunday O. Olatunji[26]	Basic personal information, family medical history, lifestyle habits, psychological and mental stress, BMI
ARDdtwo	Basic personal information, family medical history, lifestyle habits, psychological and mental stress, BMI

Therefore, we input the features selected in this study into the RF model with optimized parameters. The input features of the different models are listed in Table.3.

Comparing the AUC results, only the joint prediction was close to the accuracy of ARDdtwo. In addition, in their study, except for age, the TSH and ALB data required professional instruments for collection and testing. However, ARDdtwo has no restriction on input data and can be mined for potential correlations in any situation. Therefore, although the number of features we chose was large, they are all easily accessible and do not cause harm to human health, which improves the adaptability of the model in different application scenarios.

The AUC obtained by RF was only 0.817, indicating the superior predictive ability of our model. In a comprehensive comparison, it can be seen from Fig.7 that ARDdtwo has a significant advantage in terms of prediction effectiveness.

D. PRINCIPAL FINDINGS

The increasing annual incidence of thyroid cancer has caused widespread regulatory concern, and early diagnosis of the disease is important for early intervention and treatment. In this study on the prediction of thyroid cancer, data were obtained from The Third Xiangya Hospital, and the possible factors contributing to thyroid lesions were analyzed for each

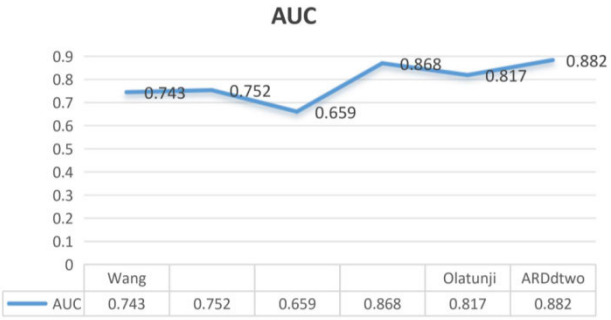


FIGURE 7. Line graph of AUC scores for each model.

age group. It is clear from the distribution of the data sample that the number of female patients was higher than that of male patients in all age groups, and the probability of the disease was approximately 2.13 times higher than that of males, which means that women are more likely to develop thyroid cancer than men. Therefore, women should be more aware of the need to prevent thyroid diseases. In addition, the number of patients in the young and old age groups was approximately 21.69% of the total number of patients, which was much less than the number of patients in the middle-aged group. These phenomena are consistent with the conclusions obtained from previous studies that have been conducted [29], [30], [31].

Thyroid cancer is associated with many factors [28], [32]. Screening results showed differences in the effect of the same element on different age groups [34], [35], which is consistent with existing studies confirming the association between age and thyroid cancer. The proportion of explicit factors was greater for young and elderly populations, whereas the influence of implicit factors was least pronounced for young people and most pronounced for middle-aged people. This means that the factors that contribute to thyroid pathology in the young and older age groups are easier to detect and prevent than those in the middle-aged group. However, both groups should pay more attention to their thyroid health if they have a family history of diabetes or hypertension, sleep too much or have an abnormal body mass index (BMI) [33], and increase the frequency of medical check-ups. The incidence of thyroid cancer is also due to the accumulation of bad habits over time, and many studies have shown an association between stress and the development of thyroid cancer [36], [37], [38]. Therefore, it is important to maintain a happy mood and an optimistic and positive attitude towards life and good lifestyle habits. In conclusion, managing one's diet and lifestyle habits and maintaining a positive and stable mood can help reduce the risk of thyroid disease.

As with most diseases, there is a lack of concern for thyroid health among people with thyroid before symptoms become apparent, and people are often less likely to go to the hospital for a detailed examination some targeted tests are harmful and expensive [15]. Thyroid cancer screening using physical examination data has been studied previously. Screening for early thyroid cancer using high-frequency ultrasound can help in the early diagnosis of thyroid cancer [27], [39], but is limited in use but does not allow for rapid screening of large populations because it requires the use of an instrument to examine everyone. Both this study and the literature [27], [40] used univariate analysis for risk factors, which allowed the model to obtain high prediction accuracy. However, other models require recalculation or substantial adjustment when the type of disease changes, whereas the model designed in this study can be flexibly transformed in different application scenarios. When Y in the association rule changes, it can be used for the prediction of different diseases to improve its usefulness in practical applications.

Meanwhile, our study fully considers implicit elements that are easily ignored in a sparse data environment instead of using fixed thresholds. When a variable threshold is used to divide each element into implicit and explicit parts, SoCIM can evaluate the element weights more accurately. The scores of each evaluation metric in Fig. 4 clearly indicate that the predictive effect of the revised association rule algorithm is improved.

V. CONCLUSION

In this study, a disease early prediction model that can adapt to spatiotemporally unbalanced data distributions was designed. It incorporates all characteristic information in the medical examination form into the analysis. Qualitative analysis of ARDcdt is used to differentiate the types of input variables and to uncover HRLFs that are at elevated risk but are easily overlooked. Based on the CIM, a quantitative analysis module, SoCIM, can adjust and optimize the relative weights of elements during model building. Utilizing big data generated by the healthcare industry to reduce costs and improve forecasting of thyroid gland disease in a simple and effective manner. The main conclusions are summarized as follows:

(1) Because ODM uses only one dimension for data information mining, ARDdtwo conducts two mining from two dimensions to analyze the features in the database more comprehensively and precisely. In addition, two-dimensional mining redesigns the conditional filtering algorithm in distinguishing the importance of risk factors which solves the 'spoofing behavior' that ODM in one dimension may produce when the data are unbalanced.
(2) The HRLF analysis module can fully consider elements that are easily ignored and account for a small percentage of the input data, thereby improving the shortcomings of previous methods that were biased when analyzing unbalanced datasets.
(3) SoCIM can scientifically measure the relative weight of each element according to the degree of influence of each element on the prediction effect of the entire system so that the importance of the elements in the calculation process of the model can be fairly assessed,

and the prediction performance of the model can be further improved.

(4) The model can be used not only for thyroid disease prediction but also for other diseases. Moreover, the results mined by the model can provide targeted advice and methods for disease prevention, improvement of daily living habits, and reduction in disease occurrence.

Future work will include using larger datasets and testing the performance of the proposed model for different diseases.

REFERENCES

[1] M. Zhi, Y. Yuan, and L. Li, "Research progress on the association between thyroid function and mood and cognitive impairment," *J. Southeast Univ. Med. Sci.*, vol. 35, pp. 612–616, Aug. 2016.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA Cancer J. Clin.*, vol. 72, pp. 7–33, Jan. 2022.

[3] J. H. Baek, "Thyroid cancer screening: How to maximize its benefits and minimize its harms," *Endocrinol. Metabolism*, vol. 38, no. 1, pp. 75–77, Feb. 2023.

[4] M. Shi, D. Nong, M. Xin, and L. Lin, "Accuracy of ultrasound diagnosis of benign and malignant thyroid nodules: A systematic review and meta-analysis," *Int. J. Clin. Pract.*, vol. 2022, pp. 1–11, Sep. 2022.

[5] F. Qi, M. Qiu, and G. Wei, "Review on ultrasonographic diagnosis of thyroid diseases based on deep learning," *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi J. Biomed. Eng. Shengwu Yixue Gongchengxue Zazhi*, vol. 40, pp. 1027–1032, Oct. 2023.

[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[7] U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. T. Said, T. M. Ghazal, and M. Ahmad, "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022.

[8] M. Alkhodari, D. K. Islayem, F. A. Alskafi, and A. H. Khandoker, "Predicting hypertensive patients with higher risk of developing vascular events using heart rate variability and machine learning," *IEEE Access*, vol. 8, pp. 192727–192739, 2020.

[9] S. Sharma and S. Bhatia, "Analysis of association rule in data mining," presented at the 2nd Int. Conf. Inf. Commun. Technol. Competitive Strategies (ICTCS), Mar. 2016, doi: 10.1145/2905055.2905238.

[10] S. Kang, L. Yang, Y. Song, R. Zhou, and J. Pang, "Satellite power system state prediction based on online learning with parameter association rules," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 3291798.

[11] G. Sheng, H. Hou, X. Jiang, and Y. Chen, "A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 695–702, Mar. 2018.

[12] C. Yuan, X. Yu, D. Li, and Y. Xi, "Overall traffic mode prediction by VOMM approach and AR mining algorithm with large-scale data," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1508–1516, Apr. 2019.

[13] A. M. Khedr, Z. A. Aghbari, A. A. Ali, and M. Eljamil, "An efficient association rule mining from distributed medical databases for predicting heart diseases," *IEEE Access*, vol. 9, pp. 15320–15333, 2021.

[14] F. Li, C. Meng, C. Wang, and S. Fan, "Data mining for risk prediction of diabetes mellitus," presented at the Int. Conf. Artif. Intell., Inf. Process. Cloud Comput. (AIIPCC), 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10070309

[15] D. Zaridze, D. Maximovitch, M. Smans, and I. Stilidi, "Thyroid cancer overdiagnosis revisited," *Cancer Epidemiol.*, vol. 74, Oct. 2021, Art. no. 102014.

[16] C.-J. Hou, R. Wei, J.-L. Tang, Q.-H. Hu, H.-F. He, and X.-M. Fan, "Diagnostic value of ultrasound features and sex of fetuses in female patients with papillary thyroid microcarcinoma," *Sci. Rep.*, vol. 8, no. 1, p. 7510, May 2018.

[17] T. Li, J. Sheng, W. Li, X. Zhang, H. Yu, X. Chen, J. Zhang, Q. Cai, Y. Shi, and Z. Liu, "A new computational model for human thyroid cancer enhances the preoperative diagnostic efficacy," *Oncotarget*, vol. 6, no. 29, pp. 28463–28477, Sep. 2015.

[18] N. M. Xi, L. Wang, and C. Yang, "Improving the diagnosis of thyroid cancer by machine learning and clinical data," *Sci. Rep.*, vol. 12, no. 1, p. 11143, Jul. 2022.

[19] R. Agrawal and R. Srikant, "Mining sequential patterns," presented at the Proc. 7th Int. Conf. Data Eng., Mar. 1995. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=380415&tag=1

[20] M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina, "Discovering symptom patterns of COVID-19 patients using association rule mining," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104249.

[21] L. I. Yujie, Z. Ruilong, and Y. Xuming, "Coronary heart disease diagnosis prediction model based on data mining technology," *Med. Inf.*, vol. 33, pp. 14–17, Nov. 2020.

[22] C. Sun, X. Wang, and Y. Zheng, "An ensemble system to predict the spatiotemporal distribution of energy security weaknesses in transmission networks," *Appl. Energy*, vol. 258, Jan. 2020, Art. no. 114062.

[23] J. Xiao, M. Liu, Q. Huang, Z. Sun, L. Ning, J. Duan, S. Zhu, J. Huang, H. Lin, and H. Yang, "Analysis and modeling of myopia-related factors based on questionnaire survey," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106162.

[24] Z. Jiang, J. Che, M. He, and F. Yuan, "A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition," *Renew. Energy*, vol. 203, pp. 802–827, Feb. 2023.

[25] R. Wang, S. Liu, Z. Hu, and W. Wu, "Study on predictors of TI-RADS grade 4a thyroid nodule nature identification," *Chongqing Med.*, vol. 51, pp. 2379–2382, Jul. 2022.

[26] S. O. Olatunji, S. Alotaibi, E. Almutairi, Z. Alrabae, Y. Almajid, R. Altabee, M. Altassan, M. I. Basheer Ahmed, M. Farooqui, and J. Alhiyafi, "Early diagnosis of thyroid cancer diseases using computational intelligence techniques: A case study of a Saudi Arabian dataset," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104267.

[27] J. Chen, M. Zhang, S. Liu, L. Ma, Y. Liu, X. Li, J. Meng, J. Li, and S. Zhang, "Thyroid cancer risk prediction based on multivariate logistic regression β value integration method for ultrasound signs," *Chin. J. Cancer*, vol. 29, pp. 289–293, Mar. 2019.

[28] Y. Li, L. Wang, J. Ni, and J. Gu, "Knowledge, awareness and perception towards thyroid cancer in general population: A systematic review," *Iranian J. Public Health*, vol. 52, pp. 219–229, Feb. 2023.

[29] P. Li, Y. Ding, M. Liu, W. Wang, and X. Li, "Sex disparities in thyroid cancer: A SEER population study," *Gland Surg.*, vol. 10, no. 12, pp. 3200–3210, Dec. 2021.

[30] Q. T. Nguyen, E. J. Lee, M. G. Huang, Y. I. Park, A. Khullar, and R. A. Plodkowski, "Diagnosis and treatment of patients with thyroid cancer," *Am Health Drug Benefits*, vol. 8, pp. 30–40, Feb. 2015.

[31] M. Kim and S. Y. Hwang, "Influence of sleep quality, coffee consumption, and perceived stress on the incidence of thyroid cancer in healthy Korean adults," *Korean J. Adult Nursing*, vol. 33, no. 2, p. 125, 2021.

[32] M. R. Youssef, A. S. C. Reisner, A. S. Attia, M. H. Hussein, M. Omar, A. LaRussa, C. A. Galvani, M. Aboueisha, M. Abdelgawad, E. A. Toraih, G. W. Randolph, and E. Kandil, "Obesity and the prevention of thyroid cancer: Impact of body mass index and weight change on developing thyroid cancer—Pooled results of 24 million cohorts," *Oral Oncol.*, vol. 112, Jan. 2021, Art. no. 105085.

[33] E. Peterson, P. De, and R. Nuttall, "BMI, diet and female reproductive factors as risks for thyroid cancer: A systematic review," *PLoS One*, vol. 7, no. 1, Jan. 2012, Art. no. e29177.

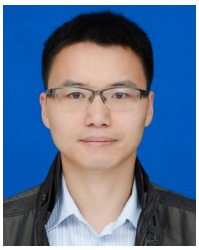
[34] N. Kwong, M. Medici, T. E. Angell, X. Liu, E. Marqusee, E. S. Cibas, J. F. Krane, J. A. Barletta, M. I. Kim, P. R. Larsen, and E. K. Alexander, "The influence of patient age on thyroid nodule formation, multinodularity, and thyroid cancer risk," *J. Clin. Endocrinol. Metabolism*, vol. 100, no. 12, pp. 4434–4440, Dec. 2015.

[35] S. Tofé, I. Argüelles, A. Forteza, C. Álvarez, A. Repetto, L. Masmiquel, I. Rodríguez, E. Losada, N. Sukunza, M. Cabrer, M. Sifontes, M. del Mar del Barrio, A. Barceló, Á. Tofé, and V. Pereg, "Age-standardized incidence, mortality rate, and trend changes of thyroid cancer in the Balearic islands during the 2000–2020 period: A population-based study," *Eur. Thyroid J.*, vol. 12, no. 3, Mar. 2023.

[36] X. Liang, Y. Zhu, and C. Kang, "Investigation of the prevalence of thyroid nodules and analysis of influencing factors," *Shenzhen J. Integrative Traditional Chin. Western Med.*, vol. 31, pp. 25–27, Apr. 2021.

[37] A. Kyriacou, V. Tziaferi, and M. Toumba, "Stress, thyroid dysregulation, and thyroid cancer in children and adolescents: Proposed impending mechanisms," *Hormone Res. Paediatrics*, vol. 96, no. 1, pp. 44–53, Mar. 2023.

[38] S. Afrashteh, M. Fararouei, M. T. Parad, and A. Mirahmadizadeh, "Sleep quality, stress and thyroid cancer: A case–control study," *J. Endocrinol. Invest.*, vol. 45, no. 6, pp. 1219–1226, Jun. 2022.
[39] Y. Wang and S. Hou, "High-frequency color ultrasound screening of early thyroid cancer in physical examination population and analysis of susceptibility factors of thyroid cancer," *Life Sci. Instrum.*, vol. 20, pp. 56–57, Jan. 2022.
[40] X. Deng, L. Tang, Y. Shen, Y. Chen, Z. Du, and Z. Zhong, "Analysis of malignant risk factors of thyroid nodules and establishment of a prediction model," *Fujian Med. J.*, vol. 40, pp. 12–15, Jun. 2018.



ZHIWEI JIA received the B.S. degree in information and computing science and the M.S. degree in operations research and cybernetics from Central South University, in 2004 and 2007, respectively, and the Ph.D. degree from the School of Electrical and Information Engineering, Shanghai Jiaotong University, in 2012. He was a Visiting Scholar with the Illinois Institute of Technology, USA, from 2017 to 2018. He has ten years of teaching and research experience. His current research interests include AI-based image classification, AI-based video target detection, and big data-based state assessment and prediction.

YUQI HUANG received the B.S. degree from the Chongqing College of Mobile Communication, in 2020. She is currently pursuing the M.S. degree in electronics and information technology with the Faculty of Electrical Engineering, Changsha University of Science and Technology, China. Her research interests include artificial intelligence, big data-based state assessment, and prediction and machine learning.

YANHUI LIN received the M.S. degree in ophthalmology from the Second Xiangya Hospital, Central South University, and the Ph.D. degree in medical science from the University of the Ryukyus, Japan. She is an attending physician in general medicine. She has presided over and participated in six national and provincial research projects and published nearly 20 SCI. Her main research interests include fundus disease and eye health management.

MIN FU graduated from the Xiangya Medical College, Central South University. She was with the Department of Ophthalmology, Zhujiang Hospital, Southern Medical University; and a Visiting Scholar with the University of California, San Francisco (UCSF). She has presided over a project of Guangdong Provincial Science and Technology Program and has participated in many national and provincial projects. She has published more than ten SCI, as the first author and the corresponding author and a number of core articles in Chinese of Peking University. Her research interests include skilled in all common and difficult diseases of ophthalmology, especially vitreoretinal diseases, specializing in various vitreoretinal surgeries and various types of lasers. She is a member of the Fundus Disease Group, Ophthalmology Branch, Guangdong Provincial Physicians Association; the Director of the Ophthalmology Branch, Chinese Famous Medicine Society; a Standing Committee Member of the Clinical Skills Training and Guidance Professional Committee, Guangdong Provincial Society of Primary Care Medicine; and also a member of the Guangdong Ophthalmic Imaging Professional Committee and the Guangdong Provincial Eye Health Care Association for the Middle-Aged and the Elderly.

CHENHAO SUN received the B.S. degree in electrical engineering from the Harbin Institute of Technology, in 2014, the M.S. degree in electrical engineering from Texas A&M University, in 2015, and the Ph.D. degree in electrical engineering from Shanghai Jiaotong University, in 2020. He is currently a Lecturer with the Changsha University of Science and Technology. His main research interests include power data mining and applications and artificial intelligence.