scientific reports

Check for updates

OPEN

A novel early diagnostic framework for chronic diseases with class imbalance

Xiaohan Yuan, Shuyu Chen✉, Chuan Sun & Lu Yuwen

Chronic diseases are one of the most severe health issues in the world, due to their terrible clinical presentations such as long onset cycle, insidious symptoms, and various complications. Recently, machine learning has become a promising technique to assist the early diagnosis of chronic diseases. However, existing works ignore the problems of feature hiding and imbalanced class distribution in chronic disease datasets. In this paper, we present a universal and efficient diagnostic framework to alleviate the above two problems for diagnosing chronic diseases timely and accurately. Specifically, we first propose a network-limited polynomial neural network (NLPNN) algorithm to efficiently capture *high-level* features hidden in chronic disease datasets, which is data augmentation in terms of its feature space and can also avoid over-fitting. Then, to alleviate the class imbalance problem, we further propose an attention-empowered NLPNN algorithm to improve the diagnostic accuracy for sick cases, which is also data augmentation in terms of its sample space. We evaluate the proposed framework on nine public and two real chronic disease datasets (partly with class imbalance). Extensive experiment results demonstrate that the proposed diagnostic algorithms outperform state-of-the-art machine learning algorithms, and can achieve superior performances in terms of accuracy, recall, F1, and G_mean. The proposed framework can help to diagnose chronic diseases timely and accurately at an early stage.

Chronic diseases have been a severe health issue in the world. In 2019, the World Health Organization pointed out that chronic diseases account for about 7 of the top 10 causes of death in the world[2]. Deaths caused by chronic diseases account for more than 63% of the total global deaths. Common chronic diseases include heart disease, diabetes, hypertension, etc., which are mainly caused by individual unhealthy lifestyles[3]. Once people suffer from chronic diseases, several vital organs (e.g., eye, brain, heart, kidney, etc.) will be damaged, and it is easy to cause a series of serious complications affecting work and life[4]. Patients with chronic diseases are particularly vulnerable to infectious diseases, such as the coronavirus disease 2019 (COVID-19)[5]. More than 48% of COVID-19 patients have a history of chronic diseases and are more likely to develop severe symptoms[6,7]. Additionally, chronic diseases will lead to expensive medical expenses[8]. The Centers for Disease Control and Prevention reports chronic diseases are leading drivers of the nation's 3.8 trillion in annual health care costs[9]. The main reason for the high fatality rate and expensive medical expenses is that chronic diseases have some terrible clinical presentations such as a long onset cycle, insidious symptoms, irreversible development, and various complications[10]. The above information reminds us that we need to quickly strengthen the prevention, diagnosis, and treatment of chronic diseases. Therefore, the early diagnosis of chronic diseases is urgent and essential, which can motivate high-risk patients to change their unhealthy lifestyles, thereby reducing the incidence of complications and further improving their health and quality of life.

Since the onset of chronic diseases is imperceptible and there are no obvious clinical symptoms in the early stage, it is difficult for doctors to determine the risk of patients with chronic diseases. Nowadays, machine learning has become the hottest promising technology for the assisted diagnosis of diseases with its advantages of autonomous learning and low error rate[11–13]. Several state-of-the-art machine learning algorithms have been widely used in the early diagnosis of different chronic diseases (e.g., chronic kidney disease, diabetes), such as support vector machines (SVM)[14], logistic regression (LR)[15], k-nearest neighbor (KNN)[16], decision trees (DT)[17], and the ensemble of some algorithms[18–20]. However, existing works are mainly dedicated to data preprocessing (e.g., data regularization and feature selection) to improve the early diagnostic performance of only a certain chronic disease[21,22]. Besides, they ignore the problems of feature hiding and imbalanced class distribution in chronic disease datasets. Hence, these methods are not conducive to improving the performance of the diagnostic model and are not suitable for a universal and efficient diagnosis of chronic diseases.

School of Big Data and Software Engineering, Chongqing University, Chongqing, China. ✉email: sychen@cqu.edu.cn

The problem of feature hiding represents that the feature in the dataset maybe not be directly related to decision-making. It needs to be further comprehensively analyzed together with other features to obtain the features directly related to decision-making[23]. For example, based on the heart rate and body mass index in the data, it is not possible to directly decide whether a patient has heart disease. If the visible original features of the data are directly used, neither the doctor nor the machine learning may be able to make a wise decision. Therefore, we need to expand the feature space of the data to capture its potential features related to chronic disease diagnosis. Additionally, the imbalanced class distribution of the dataset refers to a significant skew that exists between the number of samples for the different classes, which is also called the class imbalance problem[24]. The dominant class is called the majority class, and the remaining classes are called the minority class. Learning from the dataset with the class imbalance problem will make the learned model unreliable, which is more concerned with identifying the majority class correctly and ignoring the minority class[25,26]. Especially, in the chronic disease dataset, the number of sick cases (minority class) is generally lower than the number of healthy cases (majority class). However, the cost of misdiagnosing a sick case as a healthy case is significantly higher than the cost of misdiagnosing a healthy case as a sick case. The former may cause the patient to miss the best treatment period[27]. Therefore, how to accurately identify sick cases from the class imbalanced chronic disease dataset without affecting the overall diagnostic performance is of crucial importance and also a very challenging task.

Deep neural networks have great potential for solving various engineering problems in many fields, by extracting *high-level* features from data to achieve superior classification performance[28,29]. However, most deep neural network algorithms are not friendly to small-scale datasets and are prone to data overfitting[30,31]. Additionally, as the collected chronic disease data are not generally abundant (i.e., small-scale datasets), some existing deep neural network algorithms cannot train a well diagnostic model for chronic diseases. Recently, the deep polynomial neural network (PNN) has received the attention of some researchers[32–34]. We investigate the advantage of PNN and find that PNN is very friendly to classification tasks on small-scale datasets compared to other deep neural network algorithms. Surprisingly, the ideal PNN is parameter-free and can reduce the training error to zero iteratively[35]. Each network node of PNN is a polynomial function of its input. Thus, PNN can represent any polynomial value over the input data. Particularly, similar to other deep neural network algorithms, the network architecture of PNN is constructed layer by layer, which can represent higher and higher level (hidden) features of the input data. In other words, PNN can hierarchically expand the feature space of its input, and effectively capture features related to chronic disease diagnosis. Finally, the output layer of PNN can be constructed by solving a simple convex optimization problem.

In this paper, we are motivated to investigate the issue of the early diagnosis of chronic diseases. To the best of our knowledge, we are the first to study a universal and efficient diagnostic framework for chronic diseases, which can extract *high-level* features and solve the class imbalance problem to diagnose chronic diseases timely and accurately. Specifically, to efficiently capture *high-level* features hidden in chronic disease datasets, we propose a network-limited PNN (NLPNN) algorithm to avoid the problem of over-fitting. NLPNN can be seen as data augmentation in terms of its feature space. Additionally, as collected chronic disease datasets generally have a serious class imbalance problem, that is, the number of positive samples (sick cases) is significantly less than the number of negative samples (healthy cases), the PNN-based diagnostic model cannot fully learn the knowledge of sick cases, resulting in costly misdiagnosis (low recall). To alleviate this class imbalance problem, we further consider empowering samples with attention (i.e., weight) to change the importance of each sample and propose an improved NLPNN algorithm, named attention-empowered NLPNN (AEPNN). AEPNN pays more attention to these samples that are misclassified by NLPNN, regarded as data augmentation in terms of its sample space. Thus, the main contributions of this paper are summarized as follows.

- We study a universal and efficient diagnostic framework to make timely and accurate early diagnosis of chronic diseases with small-scale datasets.
- We propose an NLPNN algorithm to avoid the problem of over-fitting, which can efficiently capture *high-level* features hidden in chronic disease datasets and achieve high classification accuracy.
- We further propose an AEPNN algorithm to solve the class imbalance problem, which greatly improves the recall of the diagnostic model, that is, it can accurately diagnose the sick case.
- We evaluate and compare the proposed methods against other state-of-the-art methods using nine chronic diseases datasets (partly with class imbalance) and extensive experimental results demonstrate that the proposed two diagnostic models outperform state-of-the-art machine learning algorithms, and can achieve superior accuracy and recall.

The rest of the paper is organized as follows. We discuss related work in "Related work" section. "Diagnostic framework for chronic diseases" section presents the proposed algorithms, and experiment results are shown in "Experimental results" section. Finally, "Conclusion" section concludes this paper.

Related work

Early diagnosis of chronic diseases. Several existing machine learning algorithms have been proposed to diagnose a certain chronic disease[36–38]. Heydari et al.[36] compared the performance of various machine learning classification algorithms in the early diagnosis of type 2 diabetes. The simulation results showed that the performance of classification techniques depends on the nature and complexity of the dataset. Khan et al.[37] developed a chronic disease risk prediction framework. To reduce the impact of outliers, Alirezaei et al.[38] incorporated K-means clustering, SVM, and meta-heuristic algorithm to diagnose diabetes disease. However, they ignored the influence of data distribution and structural changes on model generalization performance. Under the premise of not changing the structure and distribution of data, the authors in[13] proposed a diagnostic model

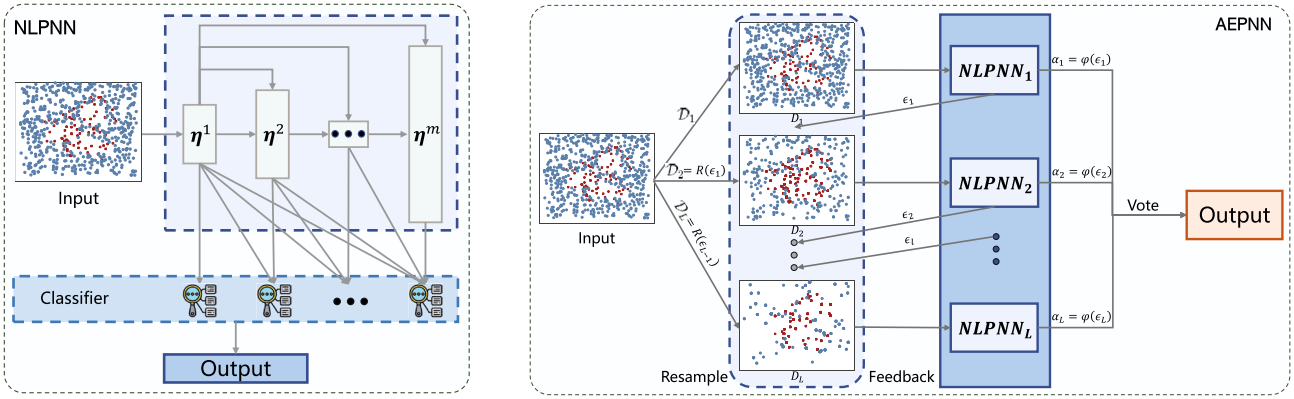


Figure 1. Flowchart of the proposed algorithms: (**a**) NLPNN; (**b**) AEPNN.

based on XGBoost for chronic kidney disease (CKD). Sekar et al.[39] used a hierarchical neural network fusion method (FHNN) for the stratified diagnosis of cardiovascular disease (CVD). However, the impact of FHNN mainly depends on the optimal choice of the sub-neural network. Some tree-based ensemble learning techniques applied to early diagnosis methods of diabetes were comprehensively studied by Tama et al.[20], and the differential performance of different classification methods was evaluated through statistical significance tests. At the same time, Altan et al.[40] also compared various machine learning algorithms for the early diagnosis of chronic obstructive pulmonary disease and proposed a deep learning model to analyze multi-channel lung sounds using statistical features of Hilbert-Huang transform, which successfully achieved high classification performance of accuracy, sensitivity, and specificity of 93.67%, 91%, and 96.33%, respectively.

Class imbalance. In medical datasets, the problem of class imbalance seriously affects the accuracy of classifiers[27,24]. In most cases, it directly leads to a high rate of misdiagnosis of the disease. This is because the class imbalance of the training data brings difficulties to the algorithm learning, and the algorithm pays more attention to the majority class[41]. However, the minority class in medical datasets (sick vs. healthy) is often more important from a data mining perspective, and it usually carries critical and useful knowledge. At present, many scholars have studied the class imbalance problem, among which there are three main methods to alleviate the class imbalance[42,43]. (1) Data-level methods: in the data preprocessing stage, re-sampling is used to reduce the size of the majority class or increase the size of the minority class (or both) to balance the training set and eliminate difference. (2) Algorithm-level methods: in the training phase, the learning algorithm is modified to be suitable for mining data with imbalanced distributions. (3) Hybrid methods: the advantages of the first two methods are combined to alleviate the adverse effects of class imbalance on the results.

Diagnostic framework for chronic diseases

Statement: I confirm that all methods were performed in accordance with the relevant guidelines and regulations.

In this section, we propose a universal and efficient diagnostic framework for diagnosing chronic diseases timely and accurately. The proposed framework consists of the NLPNN algorithm and AEPNN algorithm to alleviate the problems of feature hiding and class imbalance, respectively.

Network-limited polynomial neural network. The PNN algorithm is dedicated to learning the high-level polynomial feature representation of the data through multi-layer network architecture, and finally, output features hierarchically[32,33]. Although the PNN algorithm has been proven to run in polynomial time, it still has a limitation, that is, the depth and width of the network cannot be controlled. Its network depth and width are both adaptive, and the criterion for depth stopping is until the training error is zero[35]. In the worst case, the network depth can be infinitely deepened or the network width can be as large as the number of training samples n. This will lead to severe overfitting. Hence, we present an NLPNN algorithm for the early diagnosis of chronic diseases to avoid this issue. The structure of NLPNN is shown in Fig. 1a, and the details of the NLPNN algorithm applied to chronic diseases diagnosis be described below.

For the early diagnosis of chronic diseases, we denote the labeled training dataset as $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the set of n samples with d features; $\boldsymbol{y} = \left(y_1, y_2, \ldots, y_n\right)^T$ is a n-dimensional column vector and $y_i \in \{-1, 1\}, \forall i = 1, 2, \ldots, n$. Here, $y_i = 1$ means that the i-th sample is labeled as a sick case, and $y_i = -1$ otherwise. The M-order multivariate polynomial on the sample $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id}) \in \boldsymbol{X}$ is written as

$$\boldsymbol{p}(\boldsymbol{x}_i) = \sum_{j=0}^{M} \sum_{\boldsymbol{\alpha}^{(j)}} w_{\boldsymbol{\alpha}^{(j)}} \prod_{s=1}^{d} x_{is}^{\alpha_s^{(j)}}, \tag{1}$$

where $\boldsymbol{\alpha}^{(j)}$ is a d-dimensional vector composed of non-negative integers and $\sum_{s=1}^{d}\alpha_s^{(j)}=j$; $w_{\boldsymbol{\alpha}^{(j)}}$ is a coefficient of monomial $\prod_{s=1}^{d}x_{is}^{\alpha_s^{(j)}}$ of degree j. Represent the value of each polynomial $\boldsymbol{p}$ on n samples by linear projection

$$\boldsymbol{p}\mapsto\left(\boldsymbol{p}(\boldsymbol{x}_1),\ldots,\boldsymbol{p}(\boldsymbol{x}_n)\right)^T. \tag{2}$$

According to linear algebra, there are n polynomials $\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n$, and $\left\{\left(\boldsymbol{p}_i(\boldsymbol{x}_1),\ldots,\boldsymbol{p}_i(\boldsymbol{x}_n)\right)^T\right\}_{i=1}^{n}$ form a basis of $\mathbb{R}^n$ space. Therefore, there is a coefficient vector $\boldsymbol{v}=(v_1,\ldots,v_n)$, so that $\sum_{i=1}^{n}v_i\boldsymbol{p}_i(\boldsymbol{x}_j)=y_j,\forall y_j\in(y_1,\ldots,y_n)^T\in\mathbb{R}^n$.

The network layer of PNN is constructed by solving the basis of polynomial hierarchically, and each node calculates a linear function or weighted product over its input. We denote the j-th node of the i-th layer as $\eta_j^i(\cdot)$, which actually represents a feature (original or high-level) of the input data. For the first layer, the j-th node is the degree-1 polynomial (or linear) function $\eta_j^1(\boldsymbol{x})=[1\ \boldsymbol{x}]\boldsymbol{w}_j$, and the $\left\{\left(\eta_j^1(\boldsymbol{x}_1),\ldots,\eta_j^1(\boldsymbol{x}_n)\right)^T\right\}_{j=1}^{d+1}$ is the basis of all values obtained by a polynomial of degree 1 on the training dataset. They form the columns of matrix $F^1\in\mathbb{R}^{n\times(d+1)}$ and $F_{i,j}^1=\eta_j^1(\boldsymbol{x}_i)$. So far, a single-layer network has been constructed, and its output spans all the values obtained by the linear function on the training sample.

Generally speaking, the basis of the degree-2,3,...M polynomial is also obtained in the same trick. However, we find that the basis of the degree-M multiple polynomials is composed of $(d+1)^M$ vector elements. The scale of the basis of the polynomial increases exponentially with its degree, which will run into a computational problem.

The work in[35] indicates that any degree-m polynomial can be regraded as

$$\sum_i\boldsymbol{g}_i(\boldsymbol{x})\boldsymbol{h}_i(\boldsymbol{x})+\boldsymbol{k}(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{g}_i(\boldsymbol{x})$ and $\boldsymbol{h}_i(\boldsymbol{x})$ are degree-1 and degree-$(m-1)$ polynomials respectively; $\boldsymbol{k}(\boldsymbol{x})$ is a polynomial of degree not greater than $m-1$. Since all degree-1 polynomials are spanned by the nodes at the first layer of PNN, any degree-2 polynomial can be written as

$$\sum_i\left(\sum_j\alpha_j^{(\boldsymbol{g}_i)}\eta_j^1(\boldsymbol{x})\right)\left(\sum_r\alpha_r^{(\boldsymbol{h}_i)}\eta_r^1(\boldsymbol{x})\right)+\left(\sum_j\alpha_j^{(\boldsymbol{k})}\eta_j^1(\boldsymbol{x})\right), \tag{4}$$

where $\alpha_j^{(\boldsymbol{g}_i)},\alpha_r^{(\boldsymbol{h}_i)},\alpha_j^{(\boldsymbol{k})}$ are scalar multipliers. (4) implies that the construction of the second layer of the network is based on the first layer. The matrix $[F^1\tilde{F}^2]$ is formed by concatenating the columns of F^1, $\tilde{F}^2$, which spans all values attainable by degree-2 polynomials, and

$$\tilde{F}^2=[\left(F_1^1\circ F_1^1\right)\cdots\left(F_1^1\circ F_{|F^1|}^1\right)\cdots\left(F_{|F^1|}^1\circ F_1^1\right)\cdots\left(F_{|F^1|}^1\circ F_{|F^1|}^1\right)], \tag{5}$$

where the symbol $\circ$ indicates the Hadamard product; F_1 refers to the first column of F; $|F|$ refers to the number of columns of F. Similar to degree-1 polynomial, the column subset F^2 of $\tilde{F}^2$ should be found, so that the column of $[F^1F^2]$ are the basis of column of $[F^1\tilde{F}^2]$. The second layer of the PNN is constructed by the column of F^2, which is the product of two nodes $\eta_i^1(\cdot)$ and $\eta_j^1(\cdot)$ in the first layer.

The next step is to repeat the above process. Successively, the $m=3,4,\ldots,M$ layers of the network are constructed. We represent the matrix, written as

$$\tilde{F}^m=[\left(F_1^{m-1}\circ F_1^1\right)\cdots\left(F_1^{m-1}\circ F_{|F^1|}^1\right)\cdots\left(F_{|F^{m-1}|}^{m-1}\circ F_1^1\right)\cdots\left(F_{|F^{m-1}|}^{m-1}\circ F_{|F^1|}^1\right)]. \tag{6}$$

Thus, we find a linearly independent column subset F^m of $\tilde{F}^m$, which lets the columns of matrix $[F\ F^m]$ are a basis of the columns of the augmented matrix $[F\ \tilde{F}^m]$, where the columns of $F=\left[F^1\ F^2\ \ldots\ F^{m-1}\right]$ can span the values attained by all polynomials for degree at most $m-1$ over the training dataset. In addition, it needs to be explained that the conversion of $\tilde{F}^m$ to F^m is achieved by

$$F_s^m:=W_{i(s),j(s)}F_{i(s)}^{m-1}\circ F_{j(s)}^1,s=1,\ldots,|F^m|, \tag{7}$$

where the projection matrix $W\in\mathbb{R}^{|F^{m-1}|\times|F^1|}$ and $W_{i(s),j(s)}=\sqrt{n}/\left\|\tilde{F}_s^m\right\|$. Therefore, when the M-layer network of the PNN is constructed, all the values obtained by the polynomial of degree at most M over the training dataset can be spanned by the columns of the matrix F. In fact, F stores the high-level features of the input data, the deeper layer, the higher feature.

However, for the implementation of NLPNN, we use a parameter $\Omega=(d+1,\cdots,d+1)\in\mathbb{Z}^M$ to pre-limit the depth and width of the network, which represents that the network consists of M ($|\Omega|$) non-output layers and each layer has $d+1$ nodes at most. In the first non-output layer, we use singular value decomposition on the augmented data matrix $[\mathbf{1}\ \boldsymbol{X}]$ to obtain its partial orthogonal basis, which forms the $d+1$ nodes (select the first $d+1$ main singular vectors). In the next non-output layer, a standard Orthogonal Least Squares (OLS) procedure is utilized to greedily select the partial orthogonal basis which are the first $d+1$ relevant features for diagnosis of chronic disease according to the established high-level feature set $\tilde{F}^m$. Finally, a simple linear classifier $\boldsymbol{v}_m$ with input data $F=\left[F^1\ F^2\ \ldots\ F^m\right]$ is trained. Therefore, there are M linear classifiers in the output layer. It should

be pointed out that each linear classifier $\boldsymbol{\nu}_m$ is trained by a stochastic gradient descent method, which is utilized to solve the L_2 regularization problem

$$\min_{\boldsymbol{\nu}_m,\lambda_m} \frac{1}{n}\sum_{i=1}^{n} \ell_i\left([F^1 \ \cdots \ F^m]_{i\cdot} \cdot \boldsymbol{\nu}_m, y_i\right) + \lambda_m \|\boldsymbol{\nu}_m\|_2, \tag{8}$$

where $\ell_i\left(F_{i\cdot}^m \cdot \boldsymbol{\nu}, y_i\right) = \max(0, 1 - (F_{i\cdot}^m \cdot \boldsymbol{\nu}) \cdot y_i)$ is a hinge loss and $F_{i\cdot}^m$ represents the i-th row of matrix F^m; $\lambda_m \in \Lambda$ is the regularization factor. Then combined with the value set Λ of the regularization factor, we check the network performance layer by layer on the verification dataset to find the optimal network layer and the best regularization factor. Finally, an optimal linear classifier $\boldsymbol{\nu}^*$ is obtained by

$$\min_{m} \min_{\boldsymbol{\nu}_m,\lambda_m} \frac{1}{n}\sum_{i=1}^{n} \ell_i\left([F^1 \cdots F^m]_{i\cdot} \cdot \boldsymbol{\nu}_m, y_i\right) + \lambda_m \|\boldsymbol{\nu}_m\|_2, \tag{9}$$

and the output is this optimal classifier. The purpose of NLPNN is to adaptively find features related to diagnosis from the augmented data that is augmented in terms of its feature space. The detailed process of NLPNN is shown in Algorithm 1, which briefly describes the entire process from the establishment of the network layer to the acquisition of the output layer.

Algorithm 1: NLPNN Algorithm.

Input: $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$; Ω; Λ.
Output: An optimal linear classifier $\boldsymbol{\nu}^*$.

Initialization: $F := [\,]$ and $\tilde{F}^1 := [\mathbf{1} \ \boldsymbol{X}]$;
Solve SVD of $\tilde{F}^1$: $\tilde{F}^1 = U\Sigma W^T$;
$W := [\boldsymbol{w}_1 \ \boldsymbol{w}_2 \ \cdots \ \boldsymbol{w}_{d+1}]$; $F^1 := \tilde{F}^1 W$;
for $i = 1, \cdots, |W|$ **do**
    $F_i^1 := \frac{\sqrt{n}F_i^1}{\|F_i^1\|}$; $W_i := \frac{\sqrt{n}W_i}{\|F_i^1\|}$;
end
$F := F^1$;
$(\boldsymbol{\nu}_1, \lambda_1) = \arg\min_{\boldsymbol{\nu},\lambda} \frac{1}{n}\sum_{i=1}^{n} \ell_i\left(F_{i\cdot} \cdot \boldsymbol{\nu}, y_i\right) + \lambda\|\boldsymbol{\nu}\|_2$;
for $m = 2, \ldots, |\Omega|$ **do**
    Pick a partial orthonormal basis O^F of F's columns based supervised OLS procedure;
    $\boldsymbol{y} := \boldsymbol{y} - O^F(O^F)^T\boldsymbol{y}$;
    $\tilde{F}_s^m := F_{i(s)}^{m-1} \circ F_{j(s)}^1, s = 1, \ldots, |\tilde{F}^m|$;
    $C := \tilde{F}^m - O^F(O^F)^T\tilde{F}^m$;
    $C_i = \frac{C_i}{\|F_i^1\|}$ for all $i = 1, \cdots, |C|$;
    Compute orthonormal basis $O^{\boldsymbol{y}}$ of $\boldsymbol{y}$'s columns;
    Select the first b indices $i(1), \cdots, i(b)$ from sort($\left\|(O^{\boldsymbol{y}})^T C\right\|_2 \geq 0$, 'descend'), where $b \leq \Omega(m)$;
    for $s = i(1), \cdots, i(b)$ **do**
        $W_{i(s),j(s)} = \frac{\sqrt{n}}{\|\tilde{F}_s^m\|}$; $F_s^m := W_{i(s),j(s)}\tilde{F}_s^m$;
        Compute orthonormal basis O^C of columns of $[C_{i(1)} \ C_{i(2)} \ \cdots \ C_{i(b)}]$;
        $O^F := [O^F \ O^C]$;
        $\boldsymbol{y} := \boldsymbol{y} - O^C(O^C)^T\boldsymbol{y}$;
    end
    $F := [F \ F^m]$;
    $(\boldsymbol{\nu}_m, \lambda_m) = \arg\min_{\boldsymbol{\nu},\lambda} \frac{1}{n}\sum_{i=1}^{n} \ell_i(F_{i\cdot} \cdot \boldsymbol{\nu}, y_i) + \lambda\|\boldsymbol{\nu}\|_2$;
end
$(\boldsymbol{\nu}^*, \lambda^*) := \arg\min_m \min_{\boldsymbol{\nu}_m,\lambda_m} \frac{1}{n}\sum_{i=1}^{n} \ell_i\left([F^1 \cdots F^m]_{i\cdot} \cdot \boldsymbol{\nu}_m, y_i\right) + \lambda_m\|\boldsymbol{\nu}_m\|_2$;

Attention-empowered NLPNN. Some chronic disease datasets exist the class imbalance problem, where sick cases are generally scarce compared to healthy cases. However, the correct diagnosis of the minority sick cases among all cases is vital in a healthcare system. The reason is that the cost of misdiagnosing sick cases is much higher than healthy cases, where the latter only requires further examination and the former carries a life-threatening risk. During the training phase of NLPNN, since the samples of each class in the imbalanced dataset are utilized equally, the trained model tends to bias towards the majority class and ignore the samples (sick cases) in the minority class. Thus, NLPNN does not perform well in dealing with class imbalance problems and causes serious misdiagnosis of minority sick cases. Furthermore, for the early diagnosis of chronic diseases, although we are more concerned with the accurate diagnosis of sick cases, we cannot ignore the overall diagnostic accuracy. To alleviate the class imbalance problem, we empower the cases with attention (i.e., weight) and propose an AEPNN algorithm. AEPNN pays more attention to the cases misdiagnosed by NLPNN by changing the importance of these cases. Motivated by committee-based learning[25], AEPNN trains and combines multiple complementary NLPNN to further improve the performance of NLPNN in alleviating the class imbalance problem. The structure of AEPNN is shown in Fig. 1b.

For the implementation of AEPNN, we first assign an identical initial weight $\mathcal{D}_1(\boldsymbol{x}) = \frac{1}{n}$ to each sample $\boldsymbol{x}$ in the training dataset. An NLPNN classifier h_1 is trained from the training dataset $\boldsymbol{D}_1$ with the initialized weight distribution $\mathcal{D}_1$ and h_1's error ϵ_1 is fed back to the training sample, so that the training sample's distribution is adjusted by $\mathcal{D}_2(\boldsymbol{x})$. Then, the second NLPNN classifier h_2 is trained from the training dataset $\boldsymbol{D}_2$ with the weight distribution $\mathcal{D}_2$, where the weights of samples misdiagnosed by h_1 are increased in $\mathcal{D}_2$ to make h_2 pay more attention to the samples that are misdiagnosed by h_1. This process is repeated until h_L is trained after L iterations. Finally, the predicted label is obtained through the weighted combination of all NLPNN classifiers. The main process is shown in Algorithm 2.

Algorithm 2: AEPNN Algorithm.

Input: $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$; Ω; Λ ; NLPNN algorithm; Number of iterations L.

Output: $H_L(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{l=1}^{L} \alpha_l h_l(\boldsymbol{x})\right)$.

Initialize: the sample weight distribution $\mathcal{D}_1(\boldsymbol{x}) = \frac{1}{n}$;
for $l = 1, 2, \ldots, L$ **do**
  $h_l = NLPNN\,(D, \mathcal{D}_l)$;
  $\epsilon_l = \int_{\boldsymbol{x}\sim\mathcal{D}_l} e^{-f(\boldsymbol{x})h(\boldsymbol{x})} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$;
  if $\epsilon_l > 0.5$ **then**
    break
  end
  $\alpha_l = \frac{1}{2}\ln\left(\frac{1-\epsilon_l}{\epsilon_l}\right)$;
  $\mathcal{D}_{l+1}(\boldsymbol{x}) = \frac{\mathcal{D}_l(\boldsymbol{x}) e^{-\alpha_l h_l(\boldsymbol{x}) f(\boldsymbol{x})}}{Z_l}$;
end

Specifically, we denote the true label corresponding to sample $\boldsymbol{x}$ as $f(\boldsymbol{x})$, and the predicted label obtained by the NLPNN classifier as $h(\boldsymbol{x})$. Obviously, the loss function ϵ is defined as

$$\epsilon = \int_{\boldsymbol{x}\sim\mathcal{D}} \mathbb{I}\big(f(\boldsymbol{x}) \neq h(\boldsymbol{x})\big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{10}$$

where $p(\boldsymbol{x})$ represents the probability density function of $\boldsymbol{x}$ following the data distribution $\mathcal{D}$. However, ϵ has poor mathematical properties (non-convex and non-continuous), which makes it very difficult to be solved directly. To optimize the loss function more conveniently, we select a convex and continuously differentiable exponential loss function (11) to replace the loss function (10). Lemma 1 proves that $\ell_{\exp}(h \mid \mathcal{D})$ is the consistent replacement of the loss function ϵ, which means that (11) can replace (10) to update the weight $\mathcal{D}_l(\boldsymbol{x})$ of the sample and the weight α_l of the classifier in Algorithm 2.

Lemma 1 *The consistent replacement of the loss function ϵ is the exponential loss function*

$$\ell_{\exp}(h \mid \mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}} e^{-f(\boldsymbol{x})h(\boldsymbol{x})} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{11}$$

Proof Please see Appendix 1. □

In Algorithm 2, the h_1 is obtained by applying the NLPNN classifier to the initial samples distribution $\mathcal{D}_1$. When h_l is generated based on distribution $\mathcal{D}_l$, the weight α_l of the classifier h_l is obtained iteratively by minimize the exponential loss function $\ell_{\exp}(\alpha_l h_l \mid \mathcal{D}_l)$. From Lemma 2, we know that $\alpha_l = \frac{1}{2}\ln\left(\frac{1-\epsilon_l}{\epsilon_l}\right)$ is a necessary and

Datasets	Features	Samples	Positive:Negative	Missing
CKD	24	400	1:0.6	No
PIDD	8	768	1:1.87	No
T2DM	40	5642	1:11.19	Yes
CVD	11	70000	1:1.001	No
Heart	13	1025	1:0.95	No
GDM	83	1000	1:1.13	Yes
Fra_Heart	15	4240	1:5.58	Yes
Hep	19	155	1:0.24	Yes
BCW	10	699	1:1.9	Yes
Pri_hyper	33	9091	1:1.13	No
Pri_diab	28	14,525	1:12.78	No

Table 1. The composition details of chronic disease datasets.

sufficient condition for the exponential loss function $\ell_{\exp}(\alpha_l h_l \mid \mathcal{D}_l)$ to obtain the minimum value. It means that, under the encouragement of the weight α_l, the classifier h_l can achieve the best performance on the dataset $\boldsymbol{D}_l$ with distribution $\mathcal{D}_l$.

Lemma 2 *The exponential loss function* $\ell_{\exp}(\alpha_l h_l \mid \mathcal{D}_l)$ *at* $\alpha_l^* h_l$ *obtain the minimum value, where* $\alpha_l^* = \frac{1}{2}\ln\left(\frac{1-\epsilon_l}{\epsilon_l}\right)$ *and* $\epsilon_l = \int_{\boldsymbol{x}\sim\mathcal{D}_l} \mathbb{I}(f(\boldsymbol{x}) \neq h_l(\boldsymbol{x}))p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Proof Please see Appendix 2. □

H_l is the voting result of the first l NLPNN classifier $\{h_i\}_1^l$ with weights $\{\alpha_i\}_1^l$, and its error can be corrected by the next classifier h_{l+1}. Ideally, h_{l+1} can correct all errors of H_l by minimizing the exponential loss $\ell_{\exp}(H_l + h \mid \mathcal{D})$. From Lemma 3, all errors of H_l can be corrected by the NLPNN classifier h_{l+1} which is trained based on the sample weight distribution $\mathcal{D}_{l+1}(\boldsymbol{x}) = \mathcal{D}_l(\boldsymbol{x})\frac{e^{-\alpha_l f(\boldsymbol{x})h_l(\boldsymbol{x})}}{Z_l}$, where $\frac{1}{Z_l} = \frac{\int_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{l-1}(\boldsymbol{x})}\right]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_l(\boldsymbol{x})}\right]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}$ is the normalization factor to ensure that $\mathcal{D}_{l+1}$ is a distribution.

Lemma 3 *Assume that the base classifier* h_1 *is generated based on the data distribution* $\mathcal{D}_1$*, and* α_i *is the weight of the classifier* h_i*,* $H_l(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{i=1}^{l}\alpha_i h_i(\boldsymbol{x})\right)$*,* $l = 1, 2, \ldots$*, then all the false predictions of* H_l *can be corrected through the ideal base classifier* h_{l+1}*, which is generated based on the data distribution*

$$\mathcal{D}_{l+1}(\boldsymbol{x}) = \mathcal{D}_l(\boldsymbol{x})e^{-f(\boldsymbol{x})\alpha_l h_l(\boldsymbol{x})}\frac{\int_{\boldsymbol{x}\sim\mathcal{D}}e^{-f(\boldsymbol{x})H_{l-1}(\boldsymbol{x})}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{\boldsymbol{x}\sim\mathcal{D}}e^{-f(\boldsymbol{x})H_l(\boldsymbol{x})}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}. \tag{12}$$

Proof Please see Appendix 3. □

In summary, we iteratively optimize the exponential loss function by introducing two kinds of attention ($\mathcal{D}$ and α) to achieve the superiority of AEPNN on class-imbalanced datasets.

Experimental results

Some DNN models are not suitable for classification tasks with the small-scale dataset due to the over-fitting problem. However, the PNN-based deep learning algorithm performs well for the early diagnosis of chronic diseases with the small-scale dataset, due to its unique network structure. We select five state-of-the-art machine learning algorithms as the baseline algorithms, i.e. SVM[44], LR[45], KNN[46], DT[47], and multi-layer perceptron (MLP)[48].

Chornic disease datasets. To verify the effectiveness of the proposed algorithm in the early diagnosis of chronic diseases, we select nine public and two private chronic disease datasets for experiments. Nine public chronic disease datasets (i.e., http://archive.ics.uci.edu/ml, https://www.kaggle.com/datasets) include CKD, Pima Indian diabetes dataset (PIMA), CVD, Heart Disease Dataset (Heart), Framingham Heart Disease dataset (Fra_Heart), Hepatitis dataset (Hep), Breast Cancer Wisconsin dataset (BCW) in UCI Machine Learning Repository, Type 2 Diabetes Mellitus Dataset (T2DM) and Gestational Diabetes Mellitus dataset (GDM) in the Tianchi Precision Medicine Competition. They are scarce and precious, but some of them have problems, such as small size, class imbalance, and missing value. Two private chronic disease datasets (Pri_hyper dataset and Pri_diab dataset) are collected from a district in Chongqing, China. The Pri_hyper dataset consists of the health records of hypertensive patients and healthy people. The Pri_diab dataset consists of the health records of diabetic patients and healthy people. The composition details of the selected datasets are listed by Table 1, in which

		Regularization factor (λ)				
		10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}
Depth (Δ)	2	1	2	3	4	5
	3	6	7	8	9	10
	4	11	12	13	14	15
	5	16	17	18	19	20

Table 2. The bijective relationship between (Δ, λ) and Π.

column *Datasets* is shorthand for the name of the dataset; column *Features* represents the number of features; column *Samples* represents the number of samples; column *Positive: Negative* represents the ratio of the number of positive and negative samples; column *Missing* shows whether there are missing values in the corresponding dataset. Consistently, we split each chronic disease dataset randomly into a training dataset and testing dataset with 8:2, and maintain the distribution of the class before the split. For baseline algorithms that have to process missing values and regularize data, we fill the missing values with zeros and regularize the data. The implementation of proposed algorithms does not require any other data preprocessing technology.

Evaluation measurements. For the early diagnosis of chronic diseases, the generalization performance can be estimated on the test dataset. In addition to using the area under the receiver operating characteristic curve (AUC) to evaluate the performance of the model, we also selected the following evaluation indicators to evaluate the proposed algorithm:

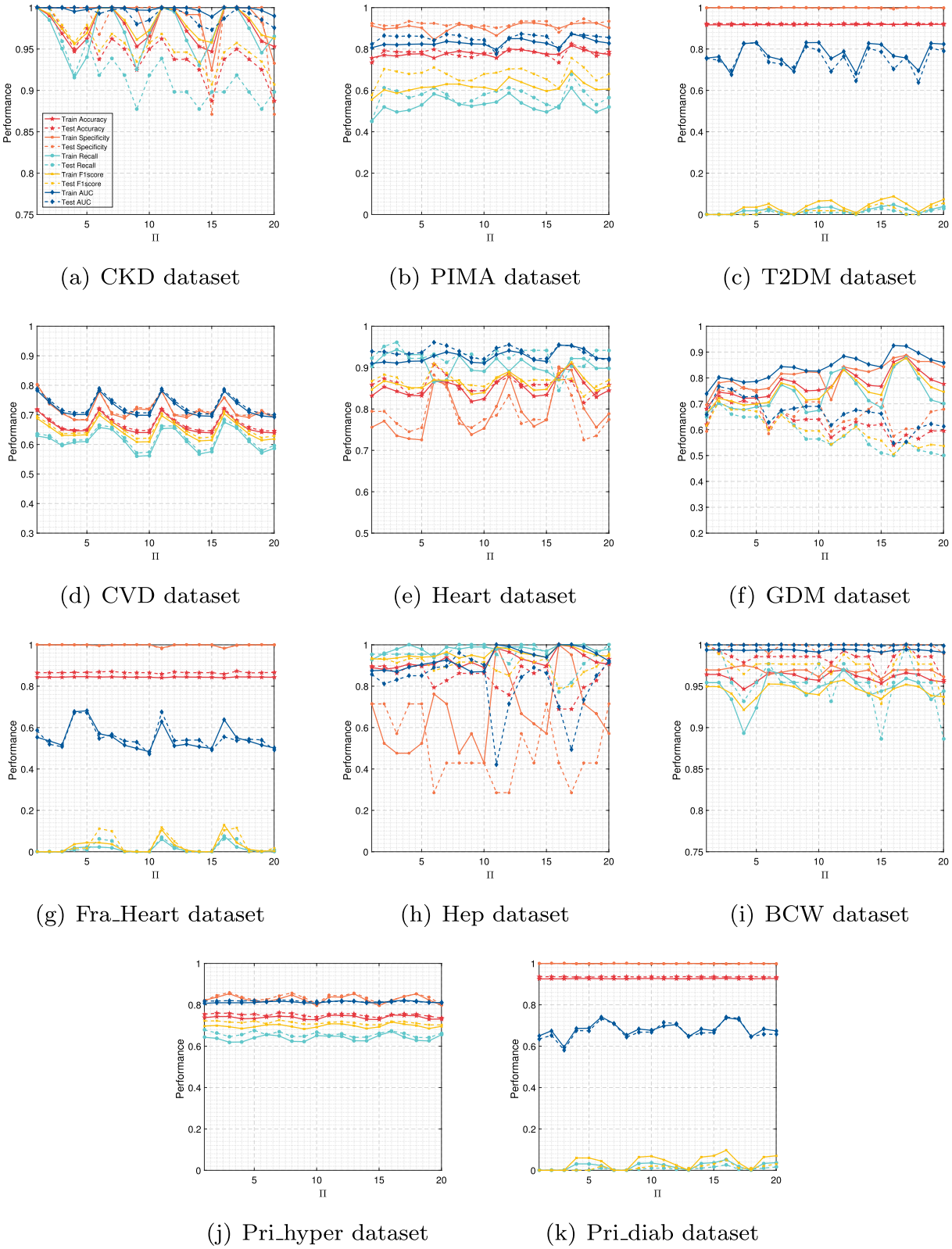
- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ represents the ratio of the number of correctly predicted specific classes to the total number of samples.
- $Specificity = \frac{TN}{TN+FP}$ represents the ratio of the number of correctly predicted healthy cases to the total healthy cases.
- $Precision = \frac{TP}{TP+FP}$ represents the ratio of the number of correctly predicted sick cases to the total predicted sick case.
- $Recall = \frac{TP}{TP+FN}$ represents the ratio of the number of correctly predicted sick cases to the total number of sick cases.
- $F1\_score = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{N+TP-TN}$ is defined based on the harmonic average of precision and recall.

where TP, FP, TN, and FN represent true positive, false positive, true negative and false negative respectively; N is the total number of samples.

Comparison of performance. We investigate the impact of different network depths Δ and regularization factor λ in the NLPNN model for the diagnostic performance of eleven chronic diseases, where $\Delta \in \{2, 3, 4, 5\}$ (network layer plus output layer) and $\lambda \in \Lambda = \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. To visually find the most suitable Δ and λ, we combine them into a binary set (Δ, λ), and establish a bijection function between (Δ, λ) and $\Pi \in \{1, 2, \cdots, 20\}$ described in Table 2. We set Π as the horizontal axis to indirectly draw the generalization performance curve of NLPNN with network depth and regularization factor. From Fig. 2, we can see that NLPNN has two advantages in the diagnosis of all chronic diseases, that is, there is no over-fitting phenomenon; the training accuracy is increasing with the increase of the number of network layers (it can be observed that when Π=1,6, 11,...). However, different Π values will affect the performance of the NLPNN algorithm, the impact on different chronic disease datasets is different.

Figure 2a shows that NLPNN can achieve 100% generalization performance on the CKD dataset when $\Pi = \{1, 2\}$. Then, with the increase of Δ and the change of λ, the test performance decreases somewhat, but both fluctuate within the range of 5%. It means that only a shallow polynomial neural network model can accurately diagnose chronic kidney disease. We can see from Fig. 2b, c, g and k that the Π value has almost no effect for the diagnostic accuracy of diabetes and heart disease. In particular, for the diagnosis of hepatitis B disease (Fig. 2h), although the accuracy of the NLPNN model does not vary greatly, its specificity is unstable with the change of Π value. This reason is that the Hep dataset has only 155 samples and the negative samples only account for 24% of the total samples. In addition, we can find the best output performance P^* of NLPNN and the corresponding value Π^* on eleven chronic disease datasets from the Fig. 2. Therefore, according to Table 2, we can find the network structure Ω^* and the regularization factor λ^* when NLPNN achieves the best performance, as shown in Table 3.

The generalization performance comparison of baseline algorithms and NLPNN algorithm on eleven chronic disease datasets are shown in Table 4, which lists the test performance results under the unified standard. In general, the diagnostic accuracy of NLPNN on the eleven chronic disease datasets is better than baseline algorithms. Especially for the diagnosis of chronic kidney disease and breast cancer, NLPNN can achieve a generalization accuracy, recall, and F1_score, of 1.0000, 1.0000, and 1.0000, respectively. In addition, NLPNN also shows significant advantages in the diagnosis of Hepatitis disease, and its generalization accuracy is about 10% better than the baseline algorithms (SVM:0.8000, LR: 0.8333, KNN: 0.8000, DT: 0.8333, MLP: 0.8000).



(a) CKD dataset (b) PIMA dataset (c) T2DM dataset

(d) CVD dataset (e) Heart dataset (f) GDM dataset

(g) Fra_Heart dataset (h) Hep dataset (i) BCW dataset

(j) Pri_hyper dataset (k) Pri_diab dataset

Figure 2. Training and test performance versus (Δ, λ) on eleven chronic disease datasets.

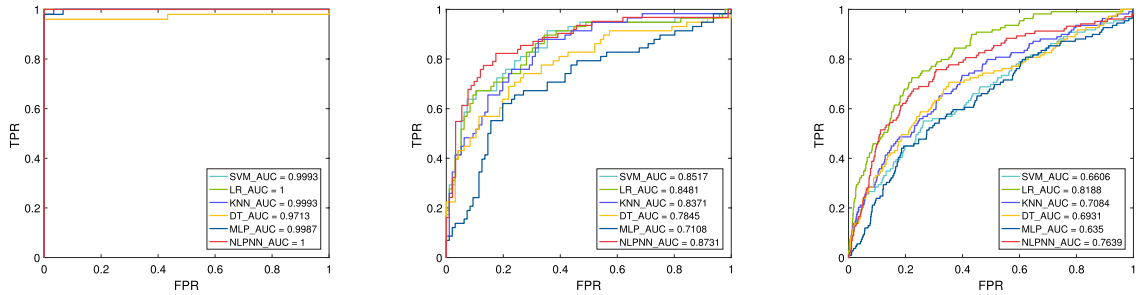
Figure 3 plots the ROC curves to further compare the performance of the NLPNN algorithm and the baseline algorithms. The AUC value of the proposed algorithm is generally better than baseline algorithms. It is also worth

		Network parameter	
		Ω^*	λ^*
Dataset	CKD	[24]	10^{-3}
	PIMA	[9 9 9 9]	10^{-2}
	T2DM	[32 32 32]	10^{-3}
	CVD	[12 12 12 12]	10^{-3}
	Heart	[13 13]	10^{-2}
	GDM	[84]	10^{-2}
	Fra_Heart	[14 14 14 14]	10^{-2}
	Hep	[14 14 14 14]	10^{0}
	BCW	[11 11 11 11]	10^{-2}
	Pri_hyper	[32 32]	10^{-2}
	Pri_diab	[29 29 29]	10^{-1}

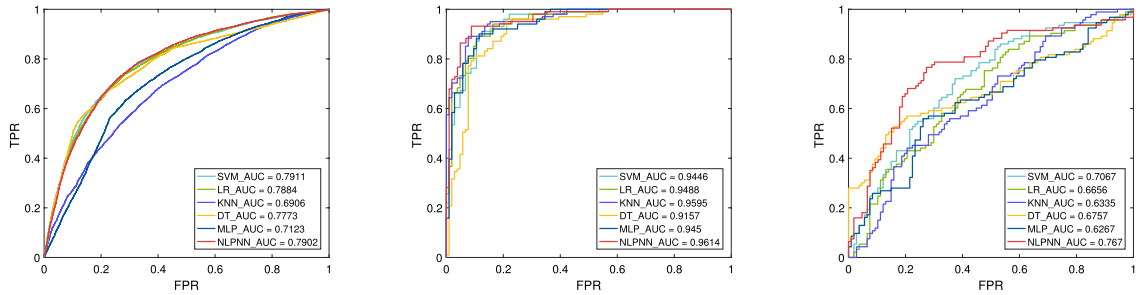
Table 3. Optimal parameter settings for different datasets.

Abbreviation		Acc	Re	F1_score	Abbreviation		Acc	Re	F1_score
CKD	SVM	0.9875	0.9800	0.9899	Fra_Heart	SVM	0.8349	0.0000	0.0000
	LR	0.9875	0.9800	0.9899		LR	0.8420	0.1000	**0.1728**
	KNN	0.9500	0.9200	0.9583		KNN	0.8337	0.0357	0.0662
	DT	0.9750	0.9600	0.9796		DT	0.8361	0.0643	0.1146
	MLP	0.9875	0.9800	0.9899		MLP	0.8314	**0.1357**	0.2099
	NLPNN	**1.0000**	**1.0000**	**1.0000**		**NLPNN**	**0.8726**	0.0614	0.1148
PIDD	SVM	0.7792	0.5517	0.6531	Hep	SVM	0.8000	1.0000	0.8846
	LR	0.7987	0.6207	0.6990		LR	0.8333	0.9565	0.8979
	KNN	0.7662	0.4827	0.6086		KNN	0.8000	0.9130	0.8749
	DT	0.7468	0.4310	0.5618		DT	0.8333	1.0000	0.9019
	MLP	0.6559	0.2414	0.3457		MLP	0.8000	0.9130	0.8749
	NLPNN	**0.8247**	**0.6774**	**0.7568**		**NLPNN**	**0.9310**	1.0000	**0.9565**
T2DM	SVM	0.9179	0.0000	0.0000	BCW	SVM	0.9714	0.9545	0.9545
	LR	0.9202	**0.0733**	**0.1311**		LR	0.9714	0.9545	0.9545
	KNN	0.9164	0.0092	0.0176		KNN	0.9714	0.9545	0.9545
	DT	0.9187	0.0092	0.0182		DT	0.9500	0.9545	0.9231
	MLP	0.9179	0.0092	0.0180		MLP	0.3143	1.0000	0.4783
	NLPNN	**0.9232**	0.0097	0.0192		**NLPNN**	**1.0000**	1.0000	**1.0000**
CVD	SVM	0.7244	0.6401	0.6977	Pri_hyper	SVM	0.7372	0.6162	0.6859
	LR	0.7249	0.6858	0.7125		LR	0.7350	0.6494	0.6953
	KNN	0.6354	0.5389	0.5950		KNN	0.7570	**0.6765**	0.7216
	DT	0.7247	0.6730	0.7085		DT	0.7224	0.4663	0.6100
	MLP	0.5385	**0.9815**	0.6789		MLP	0.7009	0.6210	0.6591
	NLPNN	**0.7265**	0.6847	**0.7161**		**NLPNN**	**0.7624**	0.6706	**0.7266**
Heart	SVM	0.8780	0.9307	0.8826	Pri_diab	SVM	0.9280	0.0000	0.0000
	LR	0.8780	0.9109	0.8804		LR	0.9315	0.0478	0.0913
	KNN	0.9024	0.8911	0.9000		KNN	0.9294	**0.1244**	**0.2023**
	DT	0.8488	0.8317	0.8442		DT	0.9325	0.0718	0.1327
	MLP	0.8878	0.8911	0.8866		MLP	0.9322	0.0861	0.1545
	NLPNN	**0.9073**	**0.9320**	**0.9100**		**NLPNN**	**0.9360**	0.0053	0.0106
GDM	SVM	0.6500	0.5591	0.5977					
	LR	0.6150	0.5376	0.5649					
	KNN	0.6200	0.4194	0.5064					
	DT	0.6950	0.5269	0.6164					
	MLP	0.6250	0.4839	0.5455					
	NLPNN	**0.7300**	**0.7021**	**0.7097**					

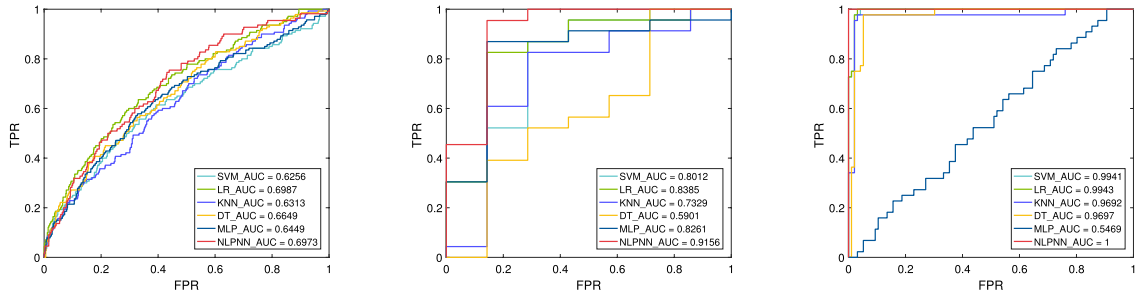
Table 4. Performance of different algorithms. The best results for each dataset are marked in bold.



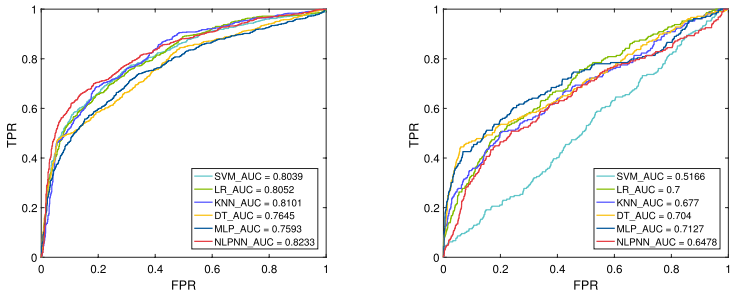
(a) Roc Curve on CKD dataset (b) Roc Curve on PIMA dataset (c) Roc Curve on T2DM dataset



(d) Roc Curve on CVD dataset (e) Roc Curve on Heart dataset (f) Roc Curve on GDM dataset



(g) Roc Curve on Fra_Heart dataset (h) Roc Curve on Hep dataset (i) Roc Curve on BCW dataset



(j) Roc Curve on Pri_hyper dataset (k) Roc Curve on Pri_diab dataset

Figure 3. ROC curves of different algorithms with the corresponding AUC values on chronic disease datasets.

noting that in the diagnosis task of chronic kidney disease and breast cancer, the NLPNN model is an "ideal model" with an AUC value of 1 (Fig. 3a, i).

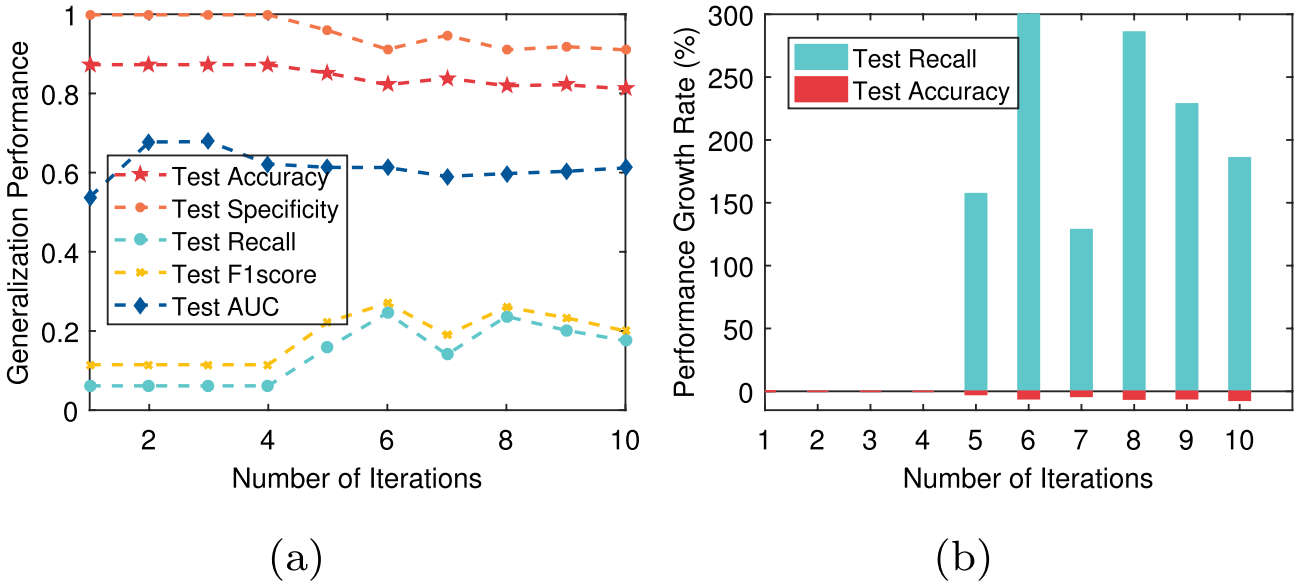


Figure 4. The test performance versus number of iteration on Fra_Heart dataset: (**a**) generalization performance; (**b**) performance growth rate.

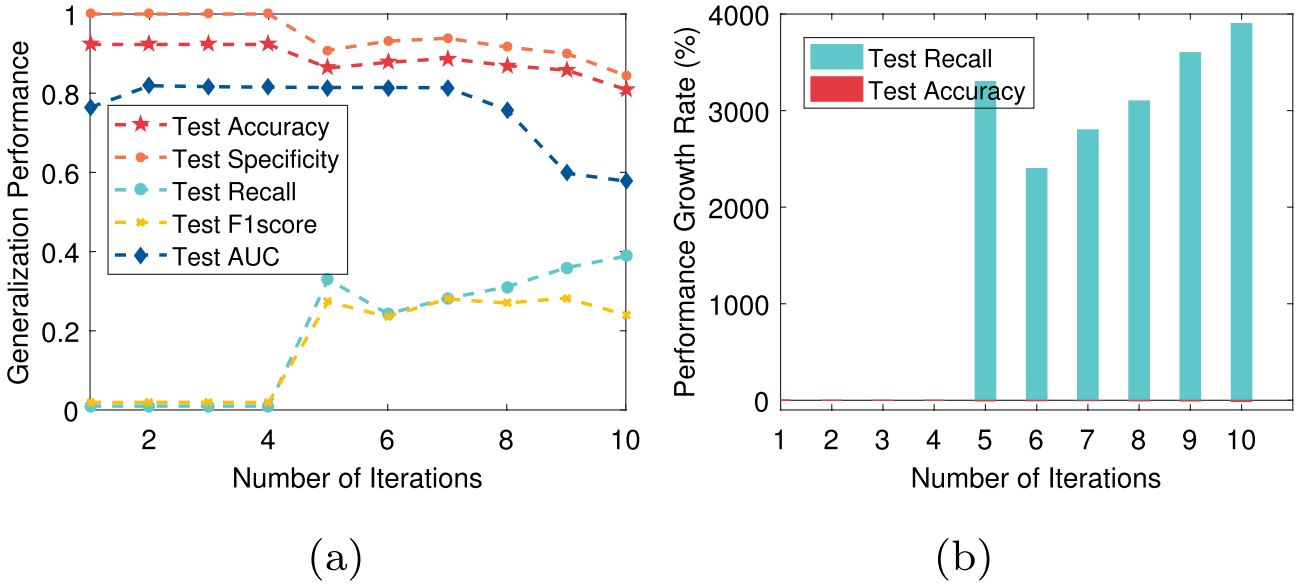


Figure 5. The test performance versus number of iteration on T2DM dataset: (**a**) generalization performance; (**b**) performance growth rate.

In this paper, we not only pay attention to the overall accuracy of the model in the diagnosis of chronic diseases but also pay more attention to whether the model can accurately diagnose sick cases (positive samples). That is, we hope that the recall of the model is as high as possible on the premise that the overall accuracy is high. For T2DM, CVD, Fra_Heart, and Pri_diab datasets, we observe that the ratio of the number of correctly predicted sick cases to the total number of sick cases is low, that is, the recall rate is low. The reason is that there is a class imbalance problem in these datasets. To solve this problem, the AEPNN algorithm 2 is proposed in "Diagnostic framework for chronic diseases" section. Because the NLPNN algorithm is a strong classifier, we do not need too many individual classifiers, whose number is equal to the number of iterations. The test performance will change with the increase of the number of training rounds of the NLPNN algorithm. Although the overall diagnostic accuracy decreases slightly, the diagnostic accuracy of sick cases has been significantly improved. We choose the number of iterations corresponding to the maximum value of the difference between the growth rate of recall and the decrease rate of accuracy as the final number of training rounds of the NLPNN algorithm to obtain the best performance. Figures 4, 5, 6, 7 show the performance of the proposed algorithm when applied to the Fra_Heart, T2DM, Pri_diab, and CVD datasets at different iterations of NLPNN, respectively. Comprehensive analysis with Table 1, we can see that the higher the class imbalance ratio of chronic disease data, the more obvious AEPNN improves the recall.

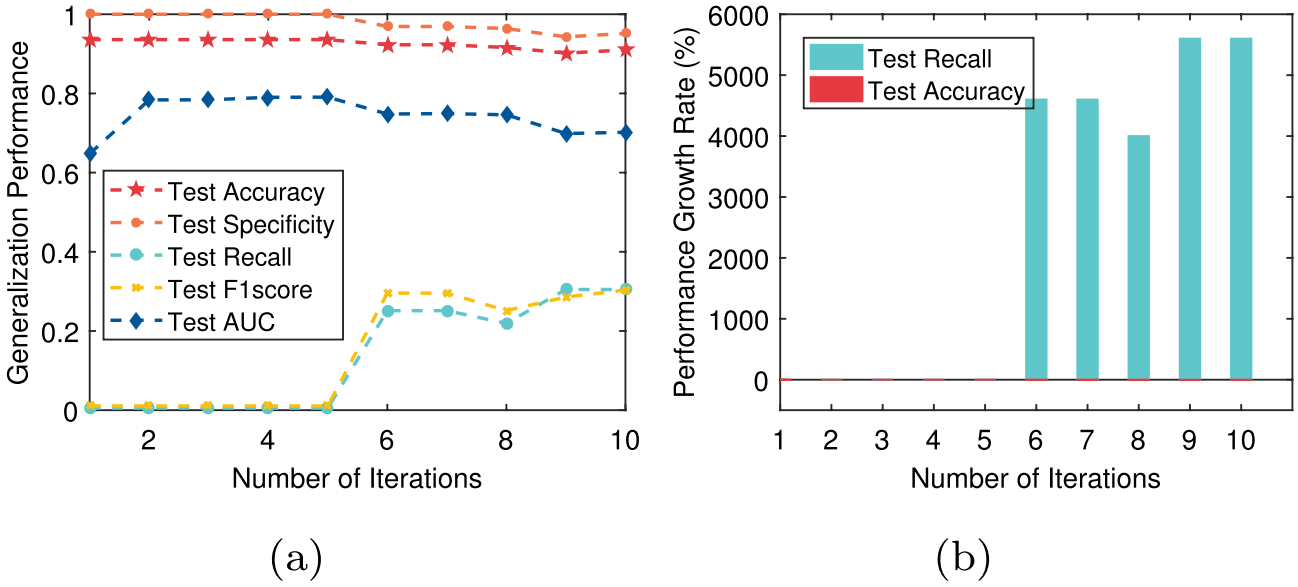


Figure 6. The test performance versus number of iteration on Pri_diab dataset: (**a**) generalization performance; (**b**) performance growth rate.

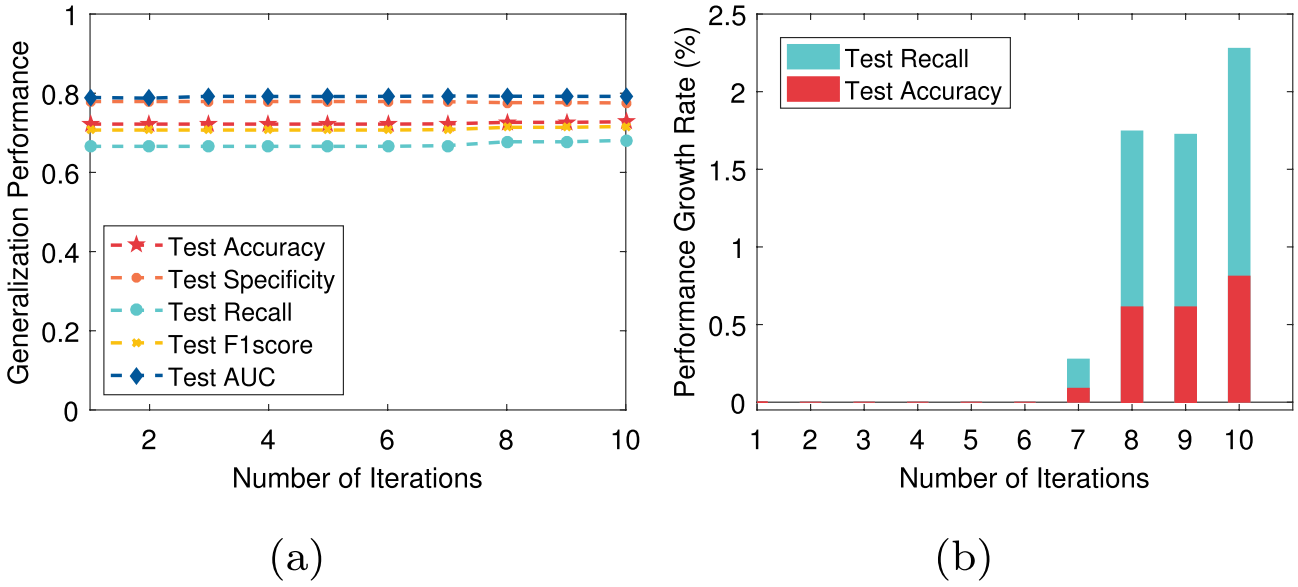


Figure 7. The test performance versus number of iteration on CVD dataset: (**a**) generalization performance; (**b**) performance growth rate.

The generalization performance of AEPNN on the Fra_Heart dataset is shown in Fig. 4a. The performance growth rate is calculated based on the number of NLPNN classifiers being one. From Fig. 4b, we observe that the recall has a growth rate of close to 300% when the number of NLPNN classifiers is six, which is chosen as the best number of NLPNN classifiers for the diagnosis of heart disease. The most surprising thing is the performance of AEPNN on the T2DM and Pri_diab datasets. As it can be seen from Figs. 5a and 6a, when the number of NLPNN is greater than four, the recall is significantly improved. When the number of NLPNN reaches ten, the growth rate of the recall approaches 4000% on the T2DM dataset and 6000% on the Pri_diab dataset. We can also know that the growth rate of recall is much higher than the decreased rate of accuracy from Figs. 5b and 6b.

From Fig. 7, we can see that although the performance of AEPNN on the CVD dataset is not significantly improved, the growth rate of recall is still higher than the decreased rate of accuracy. It indicates that the proposed algorithm is effective for the improvement of recall. The advantage it brings is that it can reduce the missed diagnosis rate for sick cases so that more patients with chronic diseases can treat and control the development of the disease in time. We also quantitatively compare the generalization performance of AEPNN and NLPNN algorithms by introducing $G\_mean = \sqrt{Recall * Specificity}$, which is a powerful indicator to evaluate the classification accuracy for class imbalanced datasets[49]. From Table 5, we can see that AEPNN can effectively improve

Abbreviation		Acc	Re	F1	G_mean
T2DM	NLPNN	0.9232	0.0097	0.0192	0.0985
	AEPNN (10/10)	0.8095	0.3883	0.2402	0.5728
CVD	NLPNN	0.7265	0.6847	0.7161	0.7256
	AEPNN (10/10)	0.7279	0.6810	0.7160	0.7267
Fra_Heart	NLPNN	0.8726	0.0614	0.1148	0.2476
	AEPNN (6/10)	0.8219	0.2456	0.2705	0.4731
Pri_diab	NLPNN	0.9360	0.0053	0.0106	0.0731
	AEPNN (10/10)	0.9098	*0.3048	0.3032	0.5385

Table 5. Performance of two proposed algorithms on four datasets with class imbalance.

G_mean by combining multiple NLPNNs. In particular, AEPNN can increase the *G_mean* from 0.0985 to 0.5728 on the T2DM dataset and from 0.0731 to 0.5385 on the Pri_diab dataset by combining ten NLPNNs.

Conclusion

In this paper, we have investigated a universal learning algorithm based on PNN for the early diagnosis of chronic diseases. Five state-of-the-art baseline algorithms are selected to compare with the NLPNN algorithm. Experiment results show that NLPNN achieves the best accuracy on the nine chronic disease datasets. In particular, for the early diagnosis of chronic kidney disease and breast cancer disease, the generalization accuracy, recall, specificity, and AUC value of this model have achieved 1.000, 1.000, 1.000, and 1.000, respectively. Furthermore, an AEPNN algorithm is further proposed to alleviate the class imbalance problem in chronic disease datasets. We aim to increase the probability of the sick cases being accurately diagnosed, that is, to increase the recall value of the model. Experiments on the four chronic disease datasets with class imbalance problems have confirmed the effectiveness of our model. It is noted that the AEPNN model performs best on the Pri_diab dataset with a positive-negative sample ratio of 1:12.78, and the growth rate of its recall is close to 6000%. The proposed algorithm can effectively assist chronic disease experts in quickly screening patients with chronic diseases, and save the cost of further testing for patients. It should be pointed out that although our algorithm performs better on small-scale datasets, the PNN-based model also shows great application potential on large-scale datasets, such as protein-protein interaction prediction and disease diagnosis based on medical images.

In future work, we will further investigate the PNN-based model in disease diagnosis. Although PNN can effectively capture hidden features parameter-free, there is still a problem with how to adaptively select the best-hidden features from the network architecture of PNN to achieve competitive performance. Thus, we consider combining PNN with computational intelligence algorithms (such as monarch butterfly optimization (MBO), earthworm optimization algorithm (EWA), and elephant herding optimization (EHO)) to improve the performance of disease diagnosis.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 18 February 2022; Accepted: 12 May 2022
Published online: 21 May 2022

References

1. Yuan, X., Chen, S., Sun, C. & Yuwen, L. A novel class imbalance-oriented polynomial neural network algorithm for disease diagnosis. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2360–2367 (2021).
2. Organization, W. H. WHO reveals leading causes of death and disability worldwide: 2000–2019. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.
3. Souza-Pereira, L., Pombo, N., Ouhbi, S., Felizardo, V. & Garcia, N. Clinical decision support systems for chronic diseases: A systematic literature review. *Comput. Methods Progr. Biomed.* **195**, 105565 (2020).
4. Alkenani, A. H., Li, Y., Xu, Y. & Zhang, Q. Predicting Alzheimer's disease from spoken and written language using fusion-based stacked generalization. *J. Biomed. Inform.* **118**, 103803 (2021).
5. Yuan, X., Chen, S., Yuwen, L., An, S., Mei, S. & Chen, T. An improved SEIR model for reconstructing the dynamic transmission of COVID-19. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2320–2327 (2020).
6. Guo, Y. *et al.* A review of wearable and unobtrusive sensing technologies for chronic disease management. *Comput. Biol. Med.* **129**, 104163 (2020).
7. Higgins, V., Sohaei, D., Diamandis, E. P. & Prassas, I. COVID-19: From an acute to chronic disease? Potential long-term health consequences. *Crit. Rev. Clin. Lab. Sci.* **58**(5), 297–310 (2021).
8. Iheanacho, I., Zhang, S., King, D., Rizzo, M. & Ismaila, A. S. Economic burden of chronic obstructive pulmonary disease (COPD): A systematic literature review. *Int. J. Chronic Obstr. Pulm. Dis.* **15**, 439 (2020).
9. For Disease Control, C., Prevention. *About Chronic Diseases*. https://www.cdc.gov/chronicdisease/about/index.htm.
10. Pathak, S. *et al.* Post-structuring radiology reports of breast cancer patients for clinical quality assurance. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**(6), 1883–1894 (2019).

11. Xia, Y., Yao, Z., Ye, Q. & Cheng, N. A dual-modal attention-enhanced deep learning network for quantification of Parkinson's disease characteristics. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**(1), 42–51. https://doi.org/10.1109/TNSRE.2019.2946194 (2020).
12. Zhang, Q., Zhou, J., Zhang, B. & Wu, E. Dsnet: Dual stack network for detecting diabetes mellitus and chronic kidney disease. *Inf. Sci.* **547**, 945–962 (2021).
13. Ogunleye, A. & Wang, Q.-G. Xgboost model for chronic kidney disease diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**(6), 2131–2140 (2019).
14. Dolatabadi, A. D., Khadem, S. E. Z. & Asl, B. M. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. *Comput. Methods Progr. Biomed.* **138**, 117–126 (2017).
15. Xiao, R., Cui, X., Qiao, H., Zheng, X. & Zhang, Y. Early diagnosis model of Alzheimer's disease based on sparse logistic regression. *Multimed. Tools Appl.* **80**(3), 3969–3980 (2021).
16. Jabbar, M. Prediction of heart disease using k-nearest neighbor and particle swarm optimization. *Biomed. Res.* **28**(9), 4154–4158 (2017).
17. Mathan, K., Kumar, P. M., Panchatcharam, P., Manogaran, G. & Varadharajan, R. A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. *Des. Automat. Embed. Syst.* **22**(3), 225–242 (2018).
18. Shang, H. & Liu, Z.-P. Prioritizing type 2 diabetes genes by weighted PageRank on bilayer heterogeneous networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**(1), 336–346 (2021).
19. Asadi, S., Roshan, S. & Kattan, M. W. Random forest swarm optimization-based for heart diseases diagnosis. *J. Biomed. Inform.* **115**, 103690 (2021).
20. Tama, B. A. & Rhee, K.-H. Tree-based classifier ensembles for early detection method of diabetes: An exploratory study. *Artif. Intell. Rev.* **51**(3), 355–370 (2019).
21. Li, J. *et al.* A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. *J. Biomed. Inform.* **115**, 103693 (2021).
22. Ma, S. *et al.* Multiple predictively equivalent risk models for handling missing data at time of prediction: With an application in severe hypoglycemia risk prediction for type 2 diabetes. *J. Biomed. Inform.* **103**, 103379 (2020).
23. Wang, G.-G., Lu, M., Dong, Y.-Q. & Zhao, X.-J. Self-adaptive extreme learning machine. *Neural Comput. Appl.* **27**(2), 291–303 (2016).
24. Singh, R. *et al.* Imbalanced breast cancer classification using transfer learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**(1), 83–93 (2020).
25. Bader-El-Den, M., Teitei, E. & Perry, T. Biased random forest for dealing with the class imbalance problem. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(7), 2163–2172 (2018).
26. Cui, Z. *et al.* Detection of malicious code variants based on deep learning. *IEEE Trans. Ind. Inform.* **14**(7), 3187–3196 (2018).
27. Yildirim, P. Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. In *Proceedings of IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, 193–198 (2017).
28. Yi, J.-H., Wang, J. & Wang, G.-G. Improved probabilistic neural networks with self-adaptive strategies for transformer fault diagnosis problem. *Adv. Mech. Eng.* **8**(1), 1687814015624832 (2016).
29. Wang, Y., Qiao, X. & Wang, G.-G. Architecture evolution of convolutional neural network using Monarch butterfly optimization. *J. Ambient Intell. Humaniz. Comput.* **13**(3), 1–15 (2022).
30. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**(1), 60–64 (2019).
31. Feng, S., Zhou, H. & Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **162**, 300–310 (2019).
32. Shi, J., Zheng, X., Li, Y., Zhang, Q. & Ying, S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.* **22**(1), 173–183 (2017).
33. Lei, H. *et al.* Protein–protein interactions prediction via multimodal deep polynomial network and regularized extreme learning machine. *IEEE J. Biomed. Health Inform.* **23**(3), 1290–1303 (2018).
34. Chrysos, G. G., Moschoglou, S., Bouritsas, G., Deng, J., Panagakis, Y. & Zafeiriou, S. P. Deep polynomial neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* https://doi.org/10.1109/TPAMI.2021.3058891 (2021).
35. Livni, R., Shalev-Shwartz, S. & Shamir, O. An algorithm for training polynomial networks. *Comput. Sci.* **26**(18), 4748–4750 (2013).
36. Heydari, M., Teimouri, M., Heshmati, Z. & Alavinia, S. M. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. *Int. J. Diabetes Dev. Ctries.* **36**(2), 167–173 (2016).
37. Khan, A., Uddin, S. & Srinivasan, U. Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes. *Expert Syst. Appl.* **136**, 230–241 (2019).
38. Alirezaei, M., Niaki, S. T. A. & Niaki, S. A. A. A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. *Expert Syst. Appl.* **127**, 47–57 (2019).
39. Sekar, B. D., Dong, M. C., Shi, J. & Hu, X. Y. Fused hierarchical neural networks for cardiovascular disease diagnosis. *IEEE Sens. J.* **12**(3), 644–650 (2011).
40. Altan, G., Kutlu, Y. & Allahverdi, N. Deep learning on computerized analysis of chronic obstructive pulmonary disease. *IEEE J. Biomed. Health Inform.* **24**(5), 1344–1350 (2019).
41. Vuttipittayamongkol, P. & Elyan, E. Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and Parkinson's disease. *Int. J. Neural Syst.* **30**(08), 2050043 (2020).
42. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Progr. Artif. Intell.* **5**(4), 221–232 (2016).
43. Sleeman, W. C. IV. & Krawczyk, B. Multi-class imbalanced big data classification on spark. *Knowl. Based Syst.* **212**, 106598 (2021).
44. Pisner, D. A. & Schnyer, D. M. Support vector machine. In *Machine Learning, Academic Press.* 101–121 (2020).
45. Nusinovici, S. *et al.* Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **122**, 56–69 (2020).
46. Khateeb, N. & Usman, M. Efficient heart disease prediction system using k-nearest neighbor classification technique. In *Proceedings of the International Conference on Big Data and Internet of Thing* 21–26 (2017).
47. Cuesta, H. A., Coffman, D. L., Branas, C. & Murphy, H. M. Using decision trees to understand the influence of individual-and neighborhood-level factors on urban diabetes and asthma. *Health Place* **58**, 102119 (2019).
48. Kirmani, M. M. Heart disease prediction using multilayer perceptron algorithm. *Int. J. Adv. Res. Comput. Sci.* **8**(5), 1169–1172 (2017).
49. Soltanzadeh, P. & Hashemzadeh, M. RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Inf. Sci.* **542**, 92–111 (2021).

Acknowledgements

Part of this work was accepted by the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA (virtually), December 9-12, 2021, which is cited as reference[1].

Author contributions

X.Y. drafted and revised the paper. X.Y. and S.C. conceived and designed the analysis and revision of this paper. X.Y. and C.S. designed and developed the framework. C.S. participated in the design and implementation of the experiment. S.C. and L.Y. have access to the dataset and performed data analysis. All the authors commented on the paper writing and the result discussion.

Funding

This work is supported by National Natural Science Foundation of China (No. 61572090), Graduate Research and Innovation Foundation of Chongqing (No. CYB21068), Chongqing Science and Technology Project (No. cstc2018jscx-mszdX0109) and the Fundamental Research Funds for the Central Universities (No. 2020CDJYGRH-YJ04).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12574-x.

Correspondence and requests for materials should be addressed to S.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.