IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 July 2023, accepted 31 August 2023, date of publication 8 September 2023, date of current version 18 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3313506

RESEARCH ARTICLE

Leveraging ISMOTE-KPCA-STACKING Algorithm for Enhanced Vascular Vertigo/Dizziness Diagnosis and Clinical Decision Support

DENGQIN SONG[1,2], TONGQIANG YI[3], QINGWEI XIANG[1,2], AND HONGCI CHEN[1,2]

[1]Geriatrics Department, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan 430061, China
[2]Clinical College of Chinese Medicine, Hubei University of Chinese Medicine, Wuhan 430065, China
[3]School of Power and Mechanical Engineering, Wuhan University, Wuhan 430072, China

Corresponding author: Hongci Chen (chc-2007@163.com)

This research was supported by the Hubei Provincial Natural Science Foundation of China (Grant No. 2022CFD022). This funding has provided us with the financial support and technical conditions necessary for our study. We express our heartfelt gratitude for this assistance.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Hubei Province Hospital of Traditional Chinese Medicine under Application No. HBZY2020-C47-01, and performed in line with the Declaration of Helsinki.

ABSTRACT Vascular vertigo/dizziness is a complex clinical syndrome involving multiple disciplines and specialties, such as neurology and psychiatry. Due to the intricate etiology and the similarity between causes and symptoms, traditional diagnostic methods based on clinical symptoms and signs are often inaccurate. This study aims to establish an effective and accurate intelligent diagnostic method for vascular dizziness to address this issue. Initially, we collected patients' medical history and biochemical indicators as research indices. To tackle the sample imbalance issue in clinical data, we employed an improved SMOTE(ISMOTE) algorithm to generate minority class data. The enhancement of the ISMOTE algorithm lies in its ability to more effectively identify and generate minority class samples in sparse regions, resolving the issue of traditional SMOTE algorithms potentially neglecting sparse areas when generating synthetic samples. Subsequently, we utilized the Pearson correlation coefficient for feature correlation analysis, screening and analyzing the original features, and identified 13 feature indices. To further improve model performance and simplify the computational process, we applied the KPCA algorithm to these indices for dimensionality reduction, ultimately obtaining three comprehensive feature indices. Finally, we constructed a Stacking ensemble algorithm model comprising base models (including KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost). To optimize the overall model performance, we introduced a fully connected cascade neural network as a meta-layer model and employed grid search and the Levenberg-Marquardt (LM) algorithm to optimize the base models and meta-layer model, respectively. This enabled the Stacking ensemble algorithm better to learn the correlations among predictions from each base model, enhancing the model's generalization ability. Experimental results demonstrate that the proposed ISMOTE-KPCA-STACKING model exhibits significant advantages in diagnosing vascular vertigo/dizziness, outperforming single base models in multiple evaluation metrics. Furthermore, the model excels in handling imbalanced data and feature selection, providing an effective method for accurately diagnosing vascular vertigo/dizziness.

INDEX TERMS Vascular vertigo/dizziness, improved SMOTE algorithm, KPCA, correlation analysis, stacking.

I. INTRODUCTION

Vertigo/dizziness is one of the most common symptoms of vertebrobasilar artery stroke, with 47% to 75% of posterior circulation stroke patients presenting with dizziness as the

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

primary symptom. In the United States, dizziness and vertigo account for 3.3% to 4.4% of emergency department visits, with stroke constituting 3% to 4% of these cases [1], [2]. Vascular vertigo/dizziness refers to central or peripheral vestibular syndrome caused by vascular diseases, which can either persist (>24hours) or be transient (<24hours). The etiology includes stroke, transient ischemic attacks, and vertebral artery compression syndrome [3]. Vascular vertigo/dizziness is classified as a malignant form of vertigo. It's important to recognize the seriousness of vascular dizziness/vertigo, as it can lead to a stroke and have a devastating impact on the patient's health. Misdiagnosing an acute stroke could have serious consequences, including missed opportunities for effective treatment and increased risk of morbidity and mortality. Conversely, overdiagnosis can lead to unnecessary examinations and treatments. Its incidence continues to rise with age [4], [5]. Typical vascular vertigo/dizziness often presents with neurological symptoms and is relatively easy to diagnose. However, when an attack does not accompany neurological symptoms, it can lead to frequent referrals and overutilization of expensive diagnostic methods, resulting in a waste of medical resources and the occurrence of adverse events. Moreover, its high incidence and recurrence rates impose a significant economic burden on patients, their families, and society, greatly diminishing patients' quality of life [6], [7].

In recent years, diagnosing vascular vertigo/dizziness has increasingly gained prominence in clinical medicine. The primary examination methods for vascular vertigo/dizziness entail medical history collection and comprehensive physical examination. As for diagnostic approaches, widely adopted methods include HINTS (Head Impulse Test, Nystagmus provoked by Gaze, Test of Skew), $ABCD^2$ score, STANDING approach, and TriAGe+ score, among others [8]. The HINTS test diagnoses by evaluating three ocular motor signs: head impulse test, nystagmus provoked by gaze, and test of skew. It has high accuracy and sensitivity but requires professional skills and equipment, potentially unsuitable for all clinical scenarios [9]. The $ABCD^2$ score, as a method of cerebral vascular risk stratification, is used to predict the risk of stroke recurrence in patients with transient ischemic attacks. Navi et al. conducted a retrospective analysis of 907 emergency patients with dizziness or vertigo. They found that the $ABCD^2$ score could identify patients at risk of cerebrovascular disease among emergency patients with dizziness and vertigo. However, its sensitivity and specificity are significantly lower than the HINTS test [10]. The STANDING approach includes differentiation between spontaneous and positional nystagmus, assessment of nystagmus direction, head impulse test, and postural balance evaluation. It can assist emergency physicians in quickly diagnosing vascular vertigo/dizziness. Vanni et al. proposed the STANDING approach for diagnosing central vertigo, with an overall accuracy of 88%, sensitivity of 95%, specificity of 87%, and negative predictive value of 99%. This method has high accuracy and reliability for non-neurotologic physicians to rule out stroke and other central vertigo cases. However, the experience and skills of the examiner may influence its accuracy [11]. In addition to the above clinical diagnostic methods, laboratory tests (such as hemodynamic assessment, lipid measurement, and blood glucose testing) and imaging techniques (such as CT, MRI, etc.) can also assist in diagnosing vascular vertigo/dizziness. In actual diagnosis, Although laboratory testing can provide valuable information, an accurate diagnosis of vascular vertigo/dizziness typically necessitates an integrated approach, combining patient history, clinical symptoms, and physical examination [8]. In imaging techniques, CT has extremely low sensitivity for acute ischemic stroke, particularly posterior circulation ischemic stroke, and its primary role in diagnosing vascular vertigo/dizziness lies in the etiological diagnosis of cerebral hemorrhage, with limited value for the etiological diagnosis of ischemic stroke [12]. Current methods for diagnosing vascular vertigo/dizziness present numerous challenges. These methods exhibit low efficiency and inadequate precision and require high technical skills from the physician, with limited applicability. These constraints result in an over-reliance on the professional expertise and experience of the healthcare provider during the diagnosis process.

With the development of computer technology, using intelligent methods to establish clinical diagnosis prediction models has become a research focus and hotspot in the medical field. Artificial intelligence, represented by machine learning, can process and analyze medical data and establish clinical diagnosis prediction models, achieving intelligent medical diagnosis and providing more efficient and effective treatment for patients [13]. In establishing clinical diagnosis prediction models, data imbalance problems are often encountered, such as the number of positive samples being far smaller than the number of negative samples. Due to the bias caused by imbalanced data, traditional machine learning algorithms often perform poorly. Researchers have conducted in-depth studies on the problem to address the data imbalance issue. Gulnaz Ahmed et al. proposed a new deep Convolutional Neural Network for detecting Alzheimer's Disease and used the ADASYN method to handle the data imbalance issue. The proposed method showed promising results in accuracy, AUC, F1 score, precision, and recall [14]. Yang et al. built a prediction model for kidney transplant rejection reactions based on SMOTE and RNN algorithms. The SMOTE algorithm reduces the imbalance between positive and negative samples and solves the problem of insufficient sample size, significantly improving model prediction accuracy [15]. Chávez-Bosquez et al. aimed to determine which oversampling algorithm could improve the performance of the Guillain-Barré syndrome (GBS) classifier by generating data for the minority class samples using three different oversampling methods (Random Over Sampling(ROS), SMOTE, and ADASYN). The results showed that the SMOTE algorithm was the best data balancing

method, which could improve the prediction model's performance [16]. Using data balancing algorithms, researchers can effectively address the issue of data imbalance, improve the accuracy and reliability of model predictions, and provide more efficient and accurate services for clinical medical diagnosis.

Features are the critical determinants of clinical diagnostic model performance, making feature selection a crucial task. Using irrelevant or redundant features decreases model performance and increases computational complexity. Researchers have conducted extensive studies and proposed various feature selection methods to select useful features and reduce unnecessary ones. For instance, Han et al. used an improved Gaussian fuzzy logic feature selection method to score feature importance and chose high-importance features to construct Alzheimer's disease classification models, resulting in good classification recognition performance [17]. Zhang et al. proposed a feature selection algorithm based on discreteness and modified correlation, increasing the correlation's impact coefficient and making feature selection more reasonable and accurate [18]. However, even with the selection of the most valuable features, practical model training processes may still involve high redundancy and correlation among features. Further feature processing is necessary, and standard methods include feature dimension reduction. Dimension reduction methods mainly include Locally Linear Embedding (LLE), Multi-dimensional Scaling (MDS), Principle Component Analysis (PCA), and so on. He et al. proposed two implementations of the quantum locally linear embedding algorithm (QLLE) for nonlinear dimensionality reduction on quantum devices, the linear-algebra-based QLLE and the variational quantum locally linear embedding algorithm (VQLLE), which achieve faster and more efficient results than the classical LLE algorithm [19]. Li et al. proposed a Density-Canopy-Kmeans clustering algorithm (DCK) for detecting network community structure, which integrates random distance and community structure coefficient based on Jaccard distance and applies Multi-dimensional Scaling (MDS) for dimension reduction. The method has demonstrated higher classification accuracy than traditional community detection methods [20]. Xu et al. developed a classification model using principal component analysis (PCA) and support vector machine (SVM) and successfully classified six types of psoriatic skin diseases. The study found that PCA, as a method for data dimensionality reduction, was effective in addressing the challenge of information overlap in classification tasks [21].

In constructing clinical diagnostic models, single classification algorithms are limited in their ability to consistently perform better for every task due to their simulation of basic data distribution. Many researchers have adopted ensemble learning algorithms for classification model construction to address these limitations. Ensemble algorithms are not a single machine learning algorithm, but instead build multiple machine learning models on a dataset and use the modeling results of all models according to specific rules as the final modeling result of the ensemble algorithm. Compared to other single algorithms, ensemble algorithms perform better and are thus widely used in the medical field. Ali, L et al. proposed the use of the XGBoost algorithm to identify and classify speech signals from Parkinson's Disease patients, and experiments have shown that its accuracy, precision, AUC, and F1-Score are superior to other algorithms, contributing to a better understanding of Parkinson's Disease and further analysis of speech features [22]. Su et al. constructed risk prediction models for coronary heart disease using three ensemble learning algorithms, balanced data categories using the SMOTE(Synthetic Minority Over-sampling Technique) algorithm, and optimized the models' hyperparameters using the Bayesian optimization algorithm. The experimental results showed that ensemble learning algorithms perform well in predicting coronary heart disease risk, with the LightGBM algorithm performing the best [23]. Nowadays, Ensemble algorithms have been widely applied to diagnosing dizziness and vertigo, especially for typical types of vertigo such as vertigo syndrome, BPPV (benign paroxysmal positional vertigo), and others. Kim et al. investigated the feasibility of utilizing the Catboost algorithm with clinical information to diagnose central vertigo. This algorithm demonstrated high accuracy, sensitivity, and specificity in diagnosing central vertigo, enabling effective classification of central vertigo based on demographic, risk factors, vital signs, and vertigo symptoms, thus providing valuable assistance for diagnosis [24]. Kamogashira et al. assessed the application of various machine learning algorithms in predicting vestibular dysfunction in vertigo patients using the center of pressure (COP) sway dataset measured during foam posturography. The results revealed that the recall of the Bagging classifier was significantly higher than logistic regression, proving that ensemble learning algorithms can successfully identify vestibular dysfunction [25]. Vascular vertigo/dizziness is a distinct type of vertigo induced by insufficient blood supply to the central nervous system. In current research on intelligent diagnostic methods, much attention has been devoted to more common types of vertigo. In contrast, diagnosing vascular vertigo/dizziness, particularly studies employing integrated algorithms and other machine learning technologies, remains relatively unexplored. This limits our capacity to understand and manage vascular vertigo/dizziness. Moreover, while vascular vertigo/dizziness typically doesn't directly threaten life, it can severely impact patients' quality of life, affecting their work capabilities, daily activities, and overall life satisfaction. Therefore, developing and optimizing intelligent diagnostic techniques for vascular vertigo/dizziness can help us diagnose this condition more accurately, offering patients more effective and personalized treatment plans.

In light of the limitations in existing research, we proposed a novel intelligent diagnostic approach for vascular

vertigo/dizziness disease. Firstly, many clinical feature indexes of both vascular vertigo/dizziness patients and normal patients were collected from medical databases and underwent in-depth analysis and selection. To balance the data distribution, an improved SMOTE algorithm was applied to generate more minority samples to improve the model's recognition performance. Regarding feature selection, we employed correlation analysis and the KPCA algorithm to screen for effective features and reduce dimensionality. Finally, a Stacking ensemble learning model based on algorithms such as KNN, RF, SVM, and GBDT was used for model training, and grid search and LM algorithm were used to optimize the base model parameters. We used evaluation methods such as recognition rate, precision, F1-score, and recall to evaluate the model's performance. Results showed that the algorithm proposed in this study had a high recognition effect. This study provides a reference paradigm for other disease studies regarding the extraction, analysis, modeling, and evaluation of patient characteristics. It offers intelligent decision-making support in the medical diagnostic process, which has important practical significance and broad application prospects.

II. MATERIALS AND METHODS

A. STUDY SUBJECTS

This study is a cross-sectional study conducted on patients with vascular vertigo/dizziness who received treatment in the Geriatric Department of Hubei Provincial Hospital of Traditional Chinese Medicine and individuals who had normal blood routine examinations during their physical examination process. As an observational research design was adopted in this study, the patient's personal information was removed, and data was analyzed anonymously with the approval of the Ethics Committee of Hubei Provincial Hospital of Traditional Chinese Medicine. In addition, the Helsinki Declaration of 1964 and its later amendments or similar ethical standards were strictly adhered to, and all participants provided written informed consent.

B. DIAGNOSTIC CRITERIA

According to the Barany Society's 2022 Diagnostic Criteria for vascular vertigo/dizziness [3]: a diagnosis of vascular vertigo/dizziness can be made when the patient presents with dizziness as the main clinical manifestation, possibly accompanied by symptoms such as nausea, vomiting, and tinnitus, and imaging studies may reveal abnormalities such as tortuous vascular courses, intracranial or extracranial arterial stenosis, and abnormal vascular development. The patient's medical history may include past events of posterior circulation dysfunction, leading to the compromised blood supply to the posterior brain hemispheres, cerebellum, inner ear vestibule, brainstem, and other areas, resulting in posterior circulation transient ischemic attack or stroke. During symptom episodes, the patient may exhibit nystagmus or other neurological signs.

C. DATA COLLECTION

1) DATA INFORMATION

In this study, we collected various clinical and biochemical indicators related to vascular vertigo. The clinical symptoms and signs encompass postischemic circulation, hypertension, hyperlipidemia, cervical spondylosis, stroke, cervical arteriosclerotic stenosis, coronary artery disease with coronary artery stenosis, and cerebral artery insufficiency. Additionally, the biochemical indicators include blood biochemistry indices for lipid metabolism, which were obtained on the day of admission. These indices consist of total cholesterol (CHOL), triglycerides (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), small dense low-density lipoprotein cholesterol (sdLDL), apolipoprotein A1 (APO-A1), apolipoprotein B (APO-B), and lipoprotein (a) (LP(a)). Detailed information on the specific indicator assignments can be found in TABLE 1.

TABLE 1. Explanation of variable assignment.

Variable	Assignment Instructions	Description	Data Type
X1	illness=1,normal=0	PCI	Discrete value
X2	illness=1,normal=0	Hypertension	Discrete value
X3	illness=1,normal=0	Hyperlipidemias	Discrete value
X4	illness=1,normal=0	Cervical Spondylosis	Discrete value
X5	illness=1,normal=0	Cerebral infarction	Discrete value
X6	illness=1,normal=0	Carotid Atherosclerosis	Discrete value
X7	illness=1,normal=0	Coronary Atherosclerosis	Discrete value
X8	illness=1,normal=0	Cerebral Circulation Insufficiency	Discrete value
I1	Measured Value	CHOL	Continuous value
I2	Measured Value	TG	Continuous value
I3	Measured Value	HDL-C	Continuous value
I4	Measured Value	LDL-C	Continuous value
I5	Measured Value	sdLDL	Continuous value
I6	Measured Value	APO-A1	Continuous value
I7	Measured Value	APO-B	Continuous value
I8	Measured Value	LP(a)	Continuous value
Y	illness=1 normal=0	vascular vertigo/dizziness	Discrete value

2) INCLUSION CRITERIA

(1) Age $\geq$ 18 years;

(2) Meet the above diagnostic criteria for vascular vertigo/dizziness;

(3) Patients and their families are aware and agree to participate in the study;

(4) Complete admission information for the patient;
(5) Biochemical indices examination conducted on the day of admission.

3) EXCLUSION CRITERIA

(1) Minors aged < 18 years;
(2) Lactating or pregnant women;
(3) Patients with serious diseases such as heart, brain, kidney, blood, tumors, etc.;
(4) Patients who have not signed the informed consent form;
(5) Patients with incomplete information;
(6) Patients who did not undergo biochemical indices examination on admission or have incomplete data.

D. DATA PREPROCESSING

Original data is often imperfect and usually contains errors and missing information. If these errors and missing information are not processed, they may affect model training results. Therefore, before inputting the dataset into the training algorithm, it is necessary to preprocess the dataset. This includes handling abnormal values, normalizing, balancing data, and extracting features.

1) ABNORMAL VALUE HANDLING

In medical data, errors in data input, measurement, and sampling can lead to the presence of outliers, which can cause deviation in experiments and produce incorrect results. Therefore, it is necessary to detect and handle outliers in the data. In this study, box plots were used to perform outlier detection. The method of identifying outliers using box plots differs from the 3σ rule, which uses statistical data such as mean and standard deviation to determine whether a data point is an outlier. The 3σ rule requires that the data is normally distributed and is only suitable for a limited range of data. Box plots, however, are a type of statistical graph that displays the overall data distribution. They do not require any specific distribution, making them well-suited for identifying outliers in real-life medical data.

For the history indicators X1-X8, there are only two possibilities, either 0 or 1, there are no outliers, so it is not considered. However, for the continuous indicators I1-I8, we need to consider the range of the indicators and remove outliers. According to the clinical indicators collected in this study, we have expanded the range of box plots to detect outliers, making them more tolerant of extreme values in the data. Data points that exceed three times the interquartile range beyond the quartiles are considered outliers. That is data points that are higher than the upper limit of the box plot (Q3 + 3IQR) or lower than the lower limit of the box plot (Q1 - 3IQR). In this study, we used box plots to analyze the I1-I8 indicators' outliers; the final results are shown in Figure 1. The red labels represent outliers, and the samples containing outliers were ultimately removed from the dataset. After the screening, we confirmed 450 vascular vertigo/dizziness samples and 150 normal samples for the final dataset.

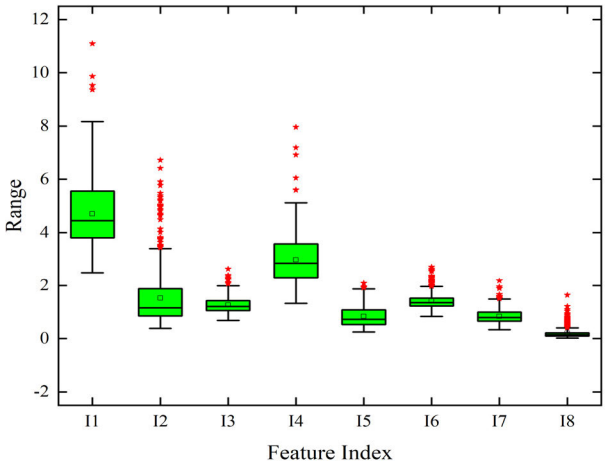

FIGURE 1. Outlier identification.

2) DATA NORMALIZATION PROCESSING

Evaluation indices often possess different dimensions and units, impacting the data analysis outcomes. To mitigate the effects of differing dimensions between indices, this paper employs the maximum value normalization method for data normalization. The data is processed using the following formula:

$$x_{normalization} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

3) DATA IMBALANCE HANDLING

Imbalanced data refers to a situation where the sample size of one or more classes in a dataset is significantly higher or lower than the sample size of other classes. The class with the larger sample size is often called the majority class, while the class with the smaller sample size is called the minority class. In imbalanced data, the information in the minority class samples may be more critical. In this study, for example, the ratio of the positive class to the negative class is approximately 3:1 (i.e., patients with vascular vertigo/dizziness to normal individuals). Based on this sample data, it is likely that a model built using this data will misdiagnose normal samples as being dizzy, which can cause significant problems and affect the model's accuracy. It is necessary to generate data to balance the data to avoid the effects of data imbalance on the model results.

SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling algorithm proposed by Chawla Kevin et al. It generates synthetic samples by interpolating between existing minority class samples based on their similarity in the feature space [26]. Specifically, it uses the k-nearest neighbors of each minority class sample as references and randomly selects N to interpolate with a threshold value within the range of [0, 1]. Its principle is as follows:

$$X_{\text{new}} = X_m + \lambda\,(X_n - X_m) \tag{2}$$

where X_{new} represents the new synthetic sample, X_m represents a minority class sample instance, X_n represents the

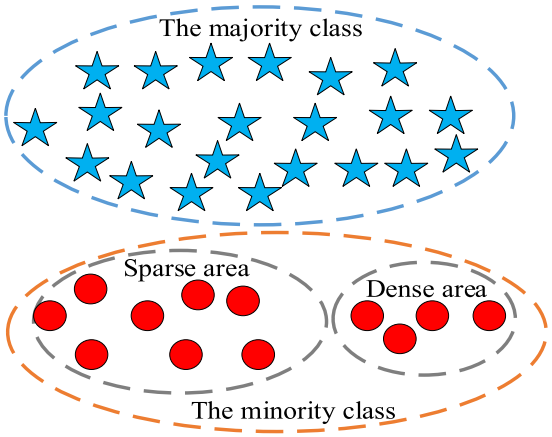

FIGURE 2. Unbalanced sample distribution.

k-nearest neighbors of X_m, and $\lambda \in [0, 1]$ is a randomly generated number.

SMOTE is a traditional algorithm that generates synthetic samples for the minority class by interpolating between existing minority class samples based on their similarity in the feature space. However, it does not consider the dataset's distribution, generating the same number of synthetic samples for each minority class sample. As illustrated in Figure 2, this can result in synthetic samples being concentrated in densely populated areas of the minority class while neglecting sparser regions that may hold crucial classification information. Consequently, this may pose challenges when training models using the synthetic samples generated by SMOTE.

Therefore, we have proposed an improved SMOTE method, which calculates the center offset value (COV) of each instance by measuring the k-neighborhood centrality as k increases, and selects the sparse samples of the minority class with high COV values for synthesis, thus increasing the diversity of the minority class.

In an n-dimensional dataset, COV is related to the minority class instance X_m and its k-neighborhood region center$C_k(X_m)$. $C_k(X_m)$ can be calculated as follows:

$$C_k(X_m) = \frac{1}{k} \sum_{q \in N_k(X_w)} \left(X_{q1}, X_{q2}, \cdots\cdots, X_{qm}\right) \tag{3}$$

where $N_k(X_w)$ is the set of k-neighbors of X_m. As the parameter k increases, we measure the displacement of $C_k(X_m)$ with $\sigma_i(X_m)$, as follows:

$$\sigma_i(X_m) = d\left(C_i(X_m), C_{i+1}(X_m)\right) \tag{4}$$

where d represents distance, $i = 1,2, \ldots\ldots, k\text{-}1$. The value of $\sigma_i(X_m)$ is typically larger in sparse areas than in dense areas. To represent the impact of k on the k-neighborhood centrality of data nodes, the absolute error of center displacement is used to represent the degree of change in the k-neighborhood centrality, and the COV coefficient is defined by accumulating the degree of change in k-neighborhood centrality as follows:

$$C_{cov}(X_m) = \sum_{i=1}^{k-2} \left|\sigma_{i+1}(X_m) - \sigma_i(X_m)\right| \tag{5}$$

Outliers located in sparse regions result in larger COV values compared to those located in dense areas. Thus, sparse samples in the minority class can be detected through large COV values.

The ISMOTE method is an effective oversampling technique for addressing data imbalance issues, specifically addressing the shortcomings of the conventional SMOTE algorithm, which inadequately considers the distribution characteristics of the dataset. By calculating the COV and measuring k-nearest neighborhood centrality for each instance, ISMOTE identifies and selects sparse minority class samples with high COV values to synthesize, thereby enhancing the diversity of the minority class. This enables ISMOTE better to capture the underlying distribution of the minority class, ultimately improving classification performance. Compared to other baseline methods like ROS and ADASYN, ISMOTE exhibits distinct advantages. The ROS method increases the proportion of the minority class by randomly replicating its samples. However, it does not generate new synthetic samples, potentially leading to model overfitting and a lack of dataset diversity. While ADASYN also focuses on generating new synthetic samples, it emphasizes creating samples based on the density distribution of minority class instances, differing from ISMOTE's approach. The primary advantage of ISMOTE over ADASYN lies in its explicit consideration of data points' k-nearest neighborhood centrality, allowing for more targeted coverage of the minority class's sparse regions when generating synthetic samples. This leads to a more diverse and representative synthetic sample collection, enhancing classification performance. Consequently, the ISMOTE method, when generating synthetic samples, considers both dataset distribution and sparse regions, resulting in significant advantages in classification performance, model generalizability, and dataset representational capacity compared to alternative oversampling methods.

4) FEATURE EXTRACTION

Before we proceed with modeling, feature extraction, and processing are crucial steps. Through feature extraction, we can better understand the information in the dataset and transform it into a form suitable for model training. In this study, we employ correlation analysis and Kernel Principal Component Analysis (KPCA) as two methods for feature processing, preparing for the subsequent model training.

1. Correlation Analysis

During the feature extraction process, correlation analysis is used to study the relationships between two variables. It helps us understand the connections between different features, guiding us in selecting features for model training. Additionally, correlation analysis assists in determining which features have the most significant impact on the model's predictions, directing us in selecting features for the final model. To better measure the correlation between

two random variables during the feature selection process, this paper uses the Pearson correlation coefficient for feature correlation analysis, aiding in understanding the relationship between features and response variables. The calculation formula is as follows:

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}} \tag{6}$$

X and Y represent two random variables, while r is the calculated correlation coefficient. The larger the absolute value of r, the stronger the correlation. When $r > 0$, the two random variables are positively correlated. When $r < 0$, the two random variables are negatively correlated. When $r = 0$, it indicates that the two random variables are uncorrelated.

2. KPCA Algorithm

Principal Component Analysis (PCA) is a dimensionality reduction technique that projects high-dimensional data onto a lower-dimensional space while preserving as much of the original variance as possible [27]. It achieves this by finding the directions in which the data varies the most, and these directions are called principal components. The first principal component accounts for the maximum variance in the data, the second main component accounts for the following highest variance, and so on. We first center the data by subtracting the mean from each feature to perform PCA. Then, we compute the covariance matrix of the centered data and calculate the eigenvectors and eigenvalues of this matrix. The eigenvectors with the highest eigenvalues are chosen as the principal components, and the data is then projected onto these components [28].

Kernel Principal Component Analysis (KPCA) is an extension of PCA that allows for non-linear dimensionality reduction. It achieves this by applying the kernel trick to the data, which maps it into a higher-dimensional feature space where the data is linearly separable. We can then use PCA in this higher-dimensional space to find the principal components and reduce the dimensionality of the data. One advantage of KPCA is that it can handle non-linearly separable data, which traditional PCA cannot. Additionally, KPCA can capture non-linear relationships in the data, which conventional PCA may miss. The mathematical formulation of KPCA can be expressed as follows:

Given a set of data points $X = \{x_1, x_2, \ldots, x_n\}$ in d-dimensional space, KPCA maps the data points into a higher-dimensional feature space using a kernel function $k(x, y)$:

$$phi\,(x) = [k\,(x, x_1)\,, k\,(x, x_2)\,, \ldots, k\,(x, x_n)] \tag{7}$$

Then, the centered kernel matrix K is calculated as:

$$K_{ij} = k(x_i, x_j) - 1/n{*}sum(k(x_i, x_j)) \tag{8}$$

The eigenvectors and eigenvalues of the matrix K are then calculated, and the eigenvectors with the highest eigenvalues are chosen as the principal components. The data is then projected onto these main components to reduce the dimensionality of the data.

As an extension of PCA, KPCA efficiently manages non-linear relationships among features. By employing a kernel function, KPCA can capture complex non-linear patterns within the data. In this study, we utilize correlation analysis and KPCA to extract meaningful features from the raw data and transform them into a form suitable for model training. These feature-processing methods not only help to reduce the computational complexity of the model but also enhance the model's predictive capabilities. Upon completing feature extraction and processing, we can input the processed features into subsequent models for training and prediction.

E. MODEL CONSTRUCTION

1) MODEL INTRODUCTION

This paper adopts the Stacking algorithm in ensemble learning to establish an accurate prediction and diagnosis model for vertigo disease. The Stacking model consists of diverse base models, including KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost, and a fully connected cascade neural network serving as a meta-learner.

1. KNN algorithm

K-Nearest Neighbors (KNN) is a straightforward instance-based algorithm for classification and regression that uses a similarity measure, such as distance, to classify new data points based on most of the closest stored cases [29]. Advantages of KNN include ease of implementation, handling of missing values, and resistance to the curse of dimensionality. However, it can be computationally expensive for large datasets, and choosing an appropriate K value is crucial for performance, which can be done using cross-validation.

2. Random Forest algorithm

Random Forest(RF)is a machine learning method for classification and regression, made up of multiple decision trees trained on randomly selected subsets of the data [30]. It predicts a sample by passing it through all decision trees and aggregating their predictions for the final result. The algorithm effectively handles missing or noisy data and has the advantage of being relatively easy to interpret. The final prediction is made by averaging the predictions of all trees, weighted by their accuracy. Random Forest has hyperparameters that can be tuned to improve performance, but it may be slower than other algorithms in large datasets.

3. The Naive Bayes algorithm

The Naive Bayes algorithm is a widely used method for classification problems. It uses the Bayes theorem to calculate the probability of a sample belonging to each class and then assigns it to the class with the highest probability [31]. It is simple to implement as it only requires calculating probabilities and making predictions based on them. Naive Bayes is also fast and handles high-dimensional data effectively. However, it makes the assumption of feature independence, which may not always hold in real-world data. Despite this limitation, Naive Bayes remains a valuable and effective

method, especially when dealing with large datasets or high feature dimensions.

4. SVM algorithm

Support Vector Machines (SVM) is a supervised learning model that can be used for both classification and regression tasks. The basic idea behind the SVM algorithm is to find a hyperplane that maximally separates the data points belonging to different classes [32]. SVM is effective in handling high-dimensional data and is also good at dealing with non-linear data. SVM is suitable for small sample size, non-linear, and high-dimensional space problems and can be used with other machine learning algorithms. SVM solves the convex quadratic optimization problem by introducing the Lagrange multiplier, which is expressed as:

$$L(\omega, \lambda, \alpha) = \frac{1}{2} \parallel \omega \parallel^2 - \sum_{i=1}^{n} \lambda_i \{[y_i (\omega x_i + b) - 1]\} \tag{9}$$

In the equation, $||\omega||$ is the norm of the normal hyperplane, b is a constant, λ_i is the Lagrange multiplier, $x_i (i = 1, 2, \ldots n)$ is the linearly separable vector, and y_i is the output class.

5. GBDT algorithm

Gradient Boosting Decision Trees (GBDT) is a machine learning algorithm that sequentially constructs decision trees for making predictions. Each successive tree is trained to rectify the errors made by its predecessors. The final prediction is determined by computing a weighted sum of all tree predictions. As an ensemble method, GBDT can manage continuous and categorical variables and address missing values. Although it is relatively resistant to overfitting, hyper-parameter choices can influence its performance, and it may be computationally demanding with large datasets. GBDT is effective for classification and regression tasks, making it well-suited for handling complex data [33].

6. XGBoost algorithm

XGBoost (eXtreme Gradient Boosting) is a machine-learning algorithm based on gradient-boosting decision trees. It is widely used in data science and machine learning competitions due to its efficiency and effectiveness [34]. XGBoost is an optimized distributed gradient boosting library that uses CART decision trees as its base classifier. It fits new functions by adding trees to predict the residuals of previous predictions, then accumulates the forecasts of all trees to obtain the final prediction result. The objective part of XGBoost is:

$$\begin{aligned} \min L &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{k} \Omega\left(f_k\right) \\ \Omega\left(f_k\right) &= \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{r} W_j^2 \end{aligned} \tag{10}$$

In the formula, n is the number of training samples, k is the number of decision trees, and f_k is the base learner. The loss function L measures the difference between the actual and predicted scores. The regularization term Ω includes two parts, where represents the number of leaf nodes and W represents the leaf node scores; γ and λ represent the penalty strength, which can control the number of leaf nodes and limit the node scores to prevent the model from overfitting to the training data and losing prediction effectiveness.

7. Fully Connected Cascade Neural Network

A fully connected cascade neural network (FCNN) is a classic deep learning model composed of multiple layers of neurons. Unlike other neural networks, each layer of neurons in an FCNN is connected to all neurons in the previous layer, meaning it is a fully connected neural network [35].

The input layer is the frontmost layer in the network and receives input data, which is transformed into output through weights W and biases b using the following formula:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \tag{11}$$

where l represents the l-th layer, $a^{(l-1)}$ represents the output of the $(l-1)$-th layer, and $z^{(l)}$ represents the input of the l-th layer.

The output layer is the final layer in the network and converts the network's output data into the final result. The output layer's output can be transformed using an activation function, with standard activation functions being the sigmoid and tanh functions.

$$a^{(l)} = \sigma(z^{(l)}) \tag{12}$$

$$a^{(l)} = tanh(z^{(l)}) \tag{13}$$

where $\sigma(z^{(l)})$ and $tanh(z^{(l)})$ represent the output of the sigmoid function and the tanh function, respectively.

The output of the network can be represented using the following formula:

$$\hat{y} = a^{(L)} \tag{14}$$

where L represents the total number of layers in the network and represents the output of the network.

8. Stacking algorithm

Ensemble learning is a technique that combines multiple models' predictions to improve overall accuracy. Stacking, a specific ensemble learning method, trains several base models and combines their predictions to form new features for training a meta-model. This method aims to enhance performance compared to any individual base model. The base models' performance and combined predictions are considered [36]. The stacking process involves dividing the feature data into training and testing sets and employing n-fold cross-validation to train the base models. As depicted in Figure 3, the training set undergoes division into n folds, where the value of n depends on the specific model employed. For each iteration, n-1 folds are utilized to train the base models, while the remaining fold is used for making predictions. This process is carried out n times, and the resulting projections are combined to generate a new feature matrix for the base models. The n-fold cross-validation ensures that the testing set is predicted n times, with the mean of these predictions taken to match the dimensions of the new feature matrix for the base models with the training set.

As shown in Figure 4 once the base models are trained and their predictions are combined into new feature matrices, they form the complete feature matrices for the meta-learner's training and testing sets. The meta-learner is then trained on these full feature matrices to make the final predictions.

In the first layer of a stacking model, diverse base models are employed to make predictions. These base models include KNN, effective for identifying patterns and predicting based on the closest data points; random forest, which handles high-dimensional data and prevents overfitting; naive Bayes, a simple and fast algorithm suitable for small datasets; and SVM, capable of managing high-dimensional and sparse data for classification and regression tasks. Additionally, GBDT, an ensemble method using decision trees as base models, excels at handling large datasets. At the same time, XGBOOST is a robust gradient-boosting algorithm frequently used in competitive machine learning competitions. Utilizing diverse base models leverages their strengths, enhancing the stacking model's overall performance. This study optimizes the first layer base models using a grid search. Grid search optimization of machine learning algorithms can be time-consuming when the range of hyperparameters is extensive. Conversely, a small hyperparameter range may result in unsatisfactory prediction performance. This study carefully selects the hyperparameter search space for each base model to address these issues. The grid search method iterates over the possible values within the chosen range. The best prediction result for each hyperparameter set is selected as the final set, aiming to balance optimization time and performance. In the second layer of the stacking model, we introduce an innovative approach by employing a fully connected cascade network with six neurons. This unique configuration enhances the model's learning capabilities, incorporating diverse activation functions in its architecture. Five of these neurons use the tanh activation function, a non-linear function that maps input values to the range (-1, 1). This activation function enables the network to capture more complex patterns within the data and adapt to a wide range of input values. The final neuron utilizes a linear sum function, aggregating the output of the previous neurons without further processing. This linear sum function serves as the network's output layer, allowing predictions based on the combined output of the preceding neurons. The fully connected cascade network, with its diverse activation functions and innovative structure, effectively learns intricate patterns in the data and results in more accurate predictions. This innovation contributes to the model's overall performance, enhancing its ability to diagnose vascular vertigo/dizziness. The Levenberg-Marquardt (LM) algorithm is employed for the second layer to optimize the fully connected cascade network weights. The LM algorithm is an iterative optimization technique that minimizes the difference between the predicted and network outputs. It starts with an initial set of weights and iteratively updates them based on the gradient of the error function, enabling the network to learn more accurate predictions. The fundamental principles of the LM algorithm are as follows:

Define the objective function $f(x)$ as the sum of squared residuals, where residuals are the differences between the predicted output and the network's output for each data point:

$$f(x) = \sum_{i=1}^{n} r_i^2 \tag{15}$$

Compute the Jacobian matrix J for the parameters x:

$$J = [\frac{\partial r_1}{\partial x_1} \cdots \frac{\partial r_1}{\partial x_n} \cdots \frac{\partial r_n}{\partial x_1} \cdots \frac{\partial r_n}{\partial x_n}] \tag{16}$$

Calculate the Hessian matrix H of the objective function $f(x)$:

$$H = \left[\frac{\partial^2 f}{\partial x_1^2} \cdots \frac{\partial^2 f}{\partial x_1 \partial x_n} \cdots \frac{\partial^2 f}{\partial x_n \partial x_1} \cdots \frac{\partial^2 f}{\partial x_n^2} \right] \tag{17}$$

Update the parameters x iteratively using the following equation:

$$x_{new} = x_{old} - \left(J^T J + \lambda I \right)^{-1} J^T \mathrm{r} \tag{18}$$

here, λ is the damping factor, and I is the identity matrix. The damping factor ensures that the Hessian matrix is positive definite, which guarantees a decrease in the objective function at each iteration.

By iteratively updating the parameters x and minimizing the objective function $f(x)$, the LM algorithm assists the fully connected cascade network make more accurate predictions for various tasks.

To summarize, the stacking ensemble model comprises two layers. The first layer consists of a diverse set of base models, including KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost. The second layer is a fully connected cascade neural network that functions as a meta-learner, combining the base models' predictions to generate final predictions. The overall model architecture is depicted in Figure 5.

2) MODEL EVALUATION INDEX

Model evaluation metrics are essential for assessing machine learning classifiers' performance and identifying improvement areas. Among these metrics, the confusion matrix, accuracy, precision, recall, and F1-score are commonly used to comprehensively understand the model's performance. Additionally, the calibration curve is employed to measure the accuracy of the classifier's predicted probabilities.

A confusion matrix is a tool used to measure the accuracy of a classifier's predictions, typically representing the counts of true positive, false positive, true negative, and false negative.

Accuracy denoted as A refers to the overall accuracy of the classifier's predictions, calculated as:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{19}$$

Precision, denoted as P, refers to the proportion of correctly predicted data by the classifier, calculated as:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{20}$$

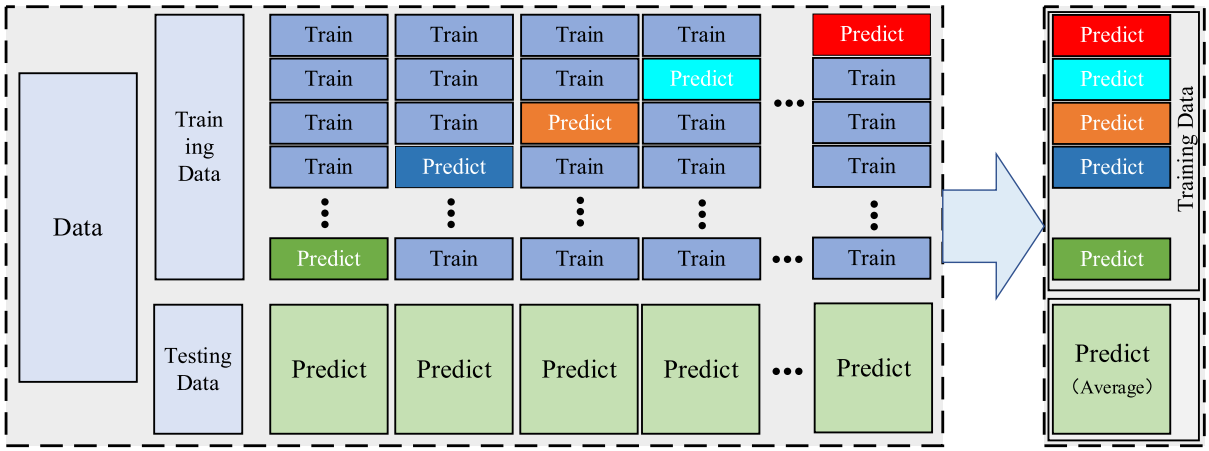

FIGURE 3. Principle of the base model.

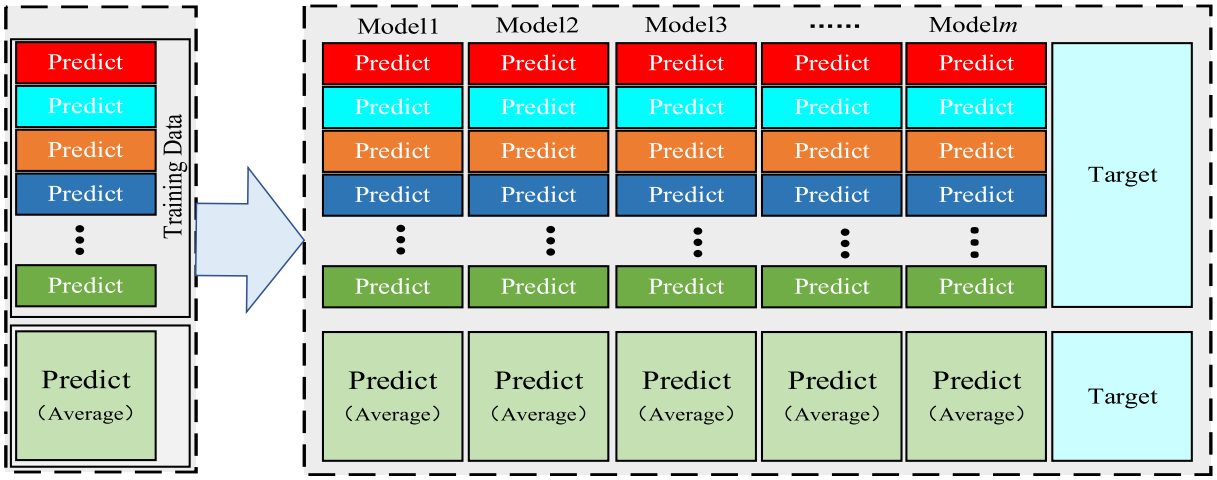

FIGURE 4. New feature matrix.

Recall, denoted as R, refers to the proportion of all correct data predicted by the classifier, calculated as:

$$R = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{21}$$

The F1-score denoted as $F1$, is the harmonic mean of precision and recall, calculated as:

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{22}$$

A calibration curve is a tool used to measure the accuracy of the classifier's predicted probabilities. The x-axis typically represents the predicted probability, and the y-axis represents the actual probability. The calibration curve will be close to the diagonal if the classifier's predicted chances are accurate. These evaluation metrics help to provide a comprehensive understanding of the model's performance and can help identify areas for improvement.

3) MODEL TRAINING

After processing the data, we train the stacking model following these steps:

1. Data Splitting

Divide the dataset into a 6:4 ratio for training and testing sets, meaning 60% of the data is allocated for training and 40% for testing. The sample distribution is detailed in TABLE 2.

TABLE 2. Data set partitioning.

Category	Training Set (Generated data)	Test Set (Generated data)	Total (Generated data)
0	273(172)	175(126)	448(298)
1	266	184	450
total	539	359	898

2. Model Construction

Build the stacking model using a Python 3 machine learning toolkit. Define the base models (KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost) and the meta-learner (FCNN).

3. Base Model Training

a. For each base model, use 5-fold cross-validation to split the training set into minor training and validation sets for evaluating model performance.

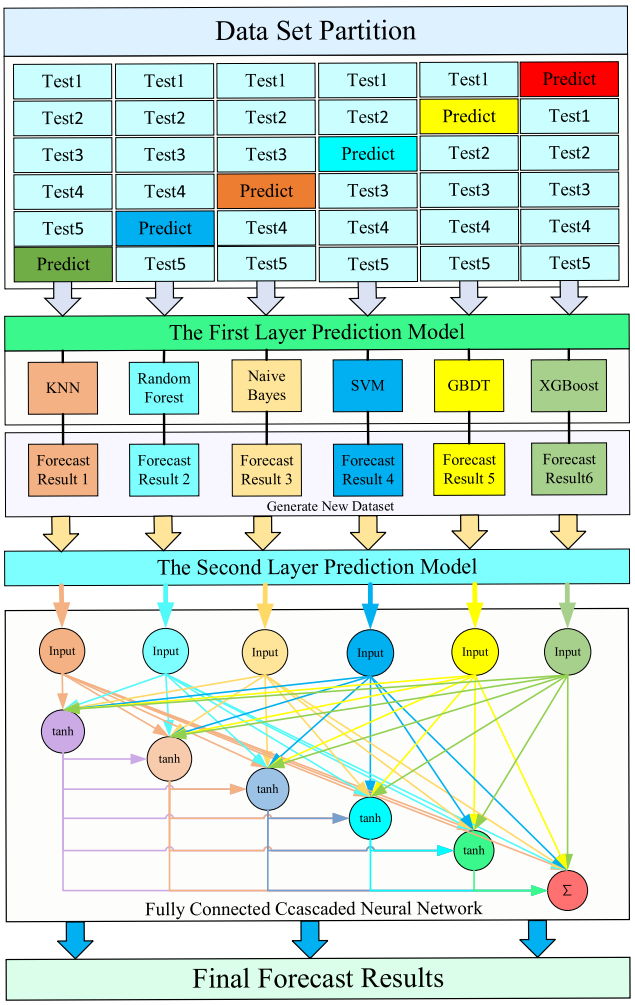

FIGURE 5. Stacking algorithm framework.

b. Train each base model on the smaller training set using their respective optimized hyperparameters.

c. Generate predictions for the validation and test sets using the trained base models.

d. Calculate the average validation set predictions for all folds of the base models.

4. Hyperparameter Optimization

Use grid search to find the optimal hyperparameters for each base model, ensuring the best performance for each model. TABLE 3 shows the optimized hyperparameters for each model.

5. Meta-learner Input Preparation

a. Combine the average validation set predictions of the base models into a new input feature set for the meta-learner. This new feature set serves as the training set for the meta-learner.

b. Combine the test set predictions of the base models into a test set for the meta-learner.

6. Meta-learner Architecture and Training

a. Use 5-fold cross-validation to split the meta-learner's training set into smaller training and validation sets.

TABLE 3. Optimized hyperparameters for base models.

Model	Hyperparameter	Optimization Range	Optimized Result
KNN	n_neighbors	1-20	12
RF	n_estimators	10-100	32
	max_depth	5-50	15
Naive Byes	alpha	0.01-2	0.5
SVM	C	0.1-10	0.1
	gamma	0.001-10	10
GBDT	n_estimators	50-200	60
	learning_rate	0.01-0.3	0.15
XGBoost	learning_rate	0.01-0.3	0.1
	gamma	0-10	10
	n_estimators	50-400	320

b. Initialize the fully connected cascade neural network with six neurons.

c. Train the meta-learner on the smaller training set using the Levenberg-Marquardt algorithm to optimize the weights. This involves calculating the Jacobian and Hessian matrices and iteratively updating the parameters to minimize the objective function.

d. Evaluate the meta-learner's performance on the validation set to monitor the training process and avoid overfitting.

e. Repeat steps 6a through 6d until all folds of the meta-learner's training set are completed.

7. Model Evaluation

a. Make predictions on the test set using the trained base models and meta-learner.

b. Calculate various performance metrics, such as accuracy, recall, F1-score, etc., to measure the model's generalization performance.

c. Compare the performance of the base models and stacking models' performance to evaluate whether the stacking approach has improved prediction accuracy.

III. RESULTS

A. FEATURE SELECTION AND ANALYSIS

We first conducted a correlation heatmap analysis on the original data containing 16 feature indices. As shown in Figure 6, I1 (TG), I4 (LDL-C), and I7 (APO-B) have high correlations, with correlation coefficients greater than 0.9. Additionally, the correlation between I3 (HDL-C) and I6 (APO-A1) is 0.85, which is also relatively high. Therefore, it is necessary to eliminate features with high correlations during the feature selection. In this study, we removed the three highly correlated indices, I4 (LDL-C), I7 (APO-B), and I6 (APO-A1), and retained the remaining 13 feature indices.

After the correlation analysis, we applied the KPCA algorithm to the remaining 13 feature indices and selected the principal components based on a cumulative contribution rate greater than 90%. As shown in Table 4, the first three principal components account for a cumulative contribution rate of 92.28%, and thus, we selected these three components. By applying the KPCA algorithm, we effectively transformed the original 13-dimensional data into three composite indices, namely the aforementioned three principal components. This methodology successfully mitigated signal overlap while

TABLE 4. The contribution rate of principal components and cumulative contribution rate.

Principal component	Contribution rate	cumulative contribution rate
1	43.75	43.75%
2	32.21	75.96%
3	16.32	92.28%
4	3.32	95.60%
5	2.26	97.86%
6	1.1	98.96%

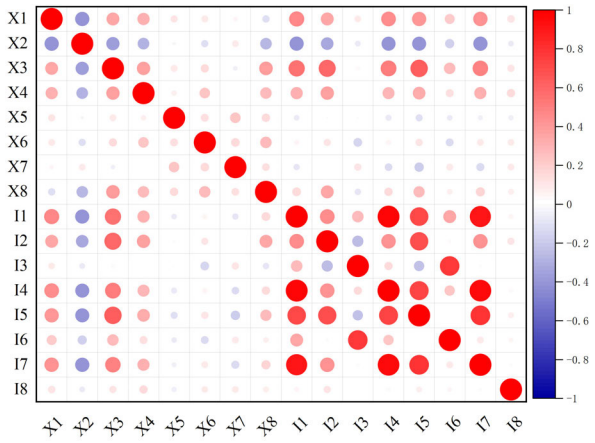

FIGURE 6. Correlation heatmap analysis.

concurrently preserving valuable information. To establish a clear connection between the data transformation and the subsequent steps in our methodology, it's crucial to mention that these three composite indices, now representing the most significant variation in data, will serve as the key features for our model training.

To demonstrate the advantages of using KPCA for feature dimensionality reduction, Figure 7 shows the dimensionality reduction results using different algorithms (MDS, LLE, PCA, and KPCA). In this figure, red denotes vertigo samples, blue represents normal samples, and yellow signifies samples generated by ISMOTE. As illustrated in Figure 7, KPCA distinctly separates and effectively clusters the different categories. Although MDS, LLE, and PCA also achieve some category separation, their clustering performance is significantly less effective than KPCA. In conclusion, KPCA is this dataset's most effective dimensionality reduction algorithm.

B. THE PROPOSED MODEL PERFORMANCE

This study employed the data above processing strategies to train and evaluate six base models (KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost) and the Stacking model. As demonstrated in Table 5, the Stacking model exhibits outstanding performance across all metrics, achieving a notable recognition rate of 99.46%. With the assistance of optimized data processing methods, the other base models also display commendable performance in accuracy, precision, recall, and F1-score, albeit with a certain degree of disparity compared to the Stacking model.

TABLE 5. Performance metrics comparison of base models and stacking model.

Model	Accuracy	Precision	Recall	F1-score
KNN	0.9722	0.9943	0.9565	0.9751
RF	0.9833	0.9837	0.9837	0.9837
NB	0.9653	0.9530	0.9892	0.9708
SVM	0.9722	0.9943	0.9565	0.9751
GBDT	0.9583	0.9829	0.9402	0.9610
XGBoost	0.9822	0.9944	0.9728	0.9835
Stacking	0.9944	0.9946	0.9946	0.9946

To better comprehend the classification performance, we employ a confusion matrix (Figure 8) to demonstrate the performance of the Stacking algorithm and six base models (KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost) on the test set. The results show that the Stacking algorithm misclassified only one diseased and one normal sample. This superior performance can be attributed to the algorithm's ability to effectively integrate the strengths of each base model, thereby enhancing the overall performance. RF and XGBoost algorithms also perform strongly, with only six misclassified samples. Notably, the RF algorithm demonstrates a more balanced classification between the two categories. While GBDT has the weakest performance among the base models, it still performs relatively well, misclassifying only 14 samples. This may be due to effective data clustering during the feature extraction process.

To further assess the prediction performance of the models, we employ a calibration curve to visualize the prediction accuracy of various models, including KNN, RF, Naive Bayes, SVM, GBDT, XGBoost, and Stacking, as the threshold varies, as shown in Figure 9. The results indicate that the Stacking algorithm outperforms the others, followed closely by RF and XGBoost algorithms. The calibration curves of these three algorithms approach the diagonal line, suggesting that their prediction results are relatively accurate. In contrast, the Naïve Bayes and GBDT algorithms demonstrate poorer performance, with their calibration curves deviating considerably from the diagonal line. The results of this study show the importance of employing a combination of optimized data processing methods, diverse base models, and ensemble algorithms to improve the performance of classification tasks in complex and imbalanced datasets.

C. COMPARISON WITH OTHER MODELS

1) THE IMPACT OF DATA BALANCING ON MODEL PERFORMANCE

Data balancing is an indispensable preprocessing step in machine learning tasks, which can enhance models' robustness and generalization performance. In this study, we assess the impact of data balancing on model performance by comparing the diagnostic accuracy of models trained on actual imbalanced data and synthetically balanced data generated by the ISMOTE algorithm. As shown in Figure 10 and Figure 11, models trained on the balanced dataset generated by the ISMOTE algorithm exhibit better diagnostic accuracy,

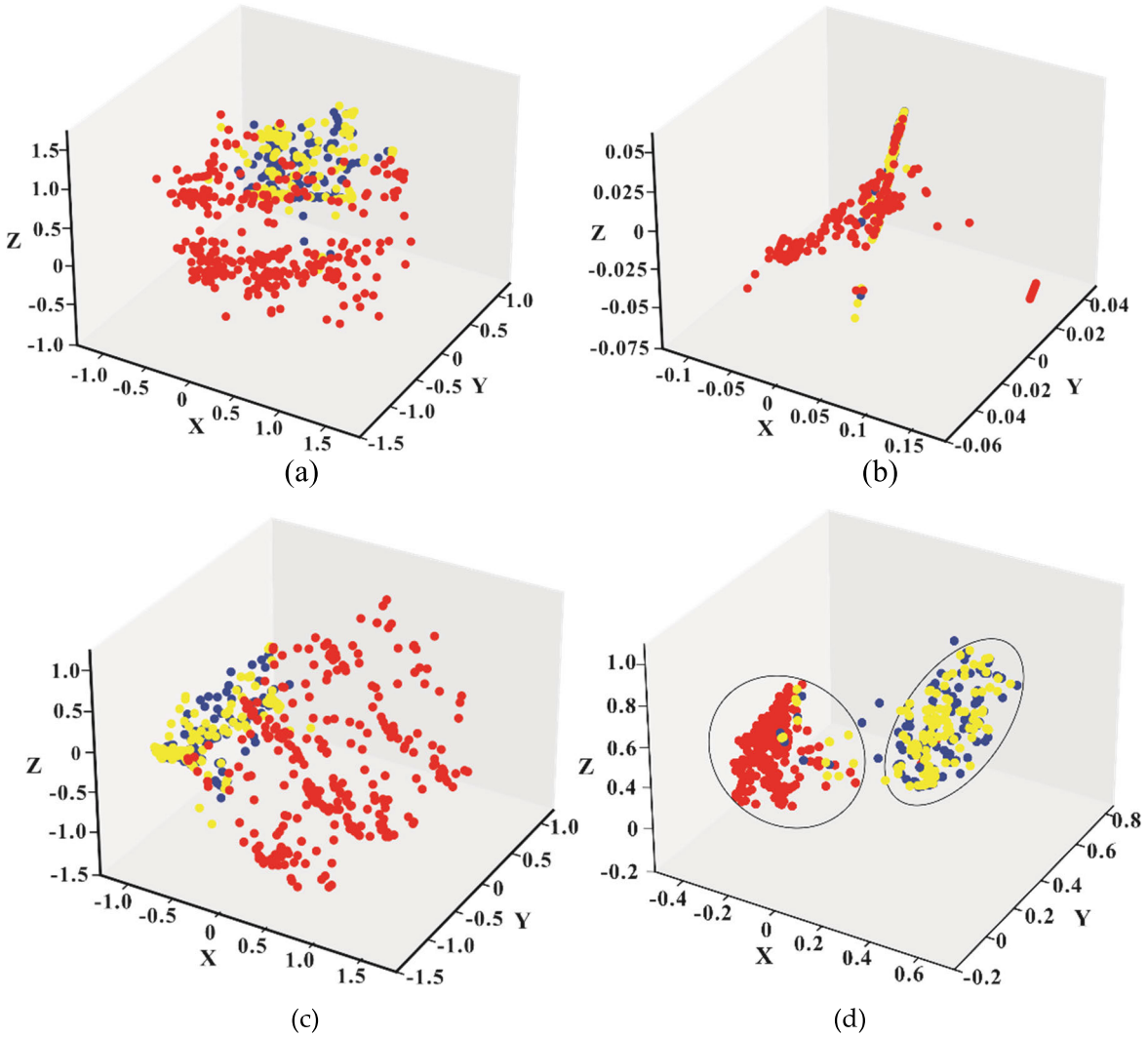

FIGURE 7. Dimensionality reduction results using different algorithms (a. MDS, b. LLE, c. PCA, d. KPCA).

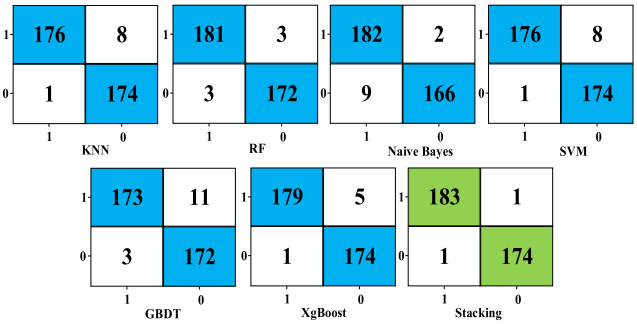

FIGURE 8. Confusion matrix for different models.

particularly in recognizing different categories more evenly. Furthermore, the Stacking model demonstrates exemplary performance on both real imbalanced data and generated balanced data, highlighting its applicability and superiority in handling imbalanced data classification tasks.

2) THE IMPACT OF DIFFERENT DATA BALANCING ALGORITHMS ON MODEL PERFORMANCE

To evaluate the advantages of the ISMOTE algorithm, we compare its performance with the ADASYN (Adaptive Synthetic Sampling) algorithm and the original SMOTE algorithm. ADASYN is an extension of the SMOTE algorithm that generates synthetic samples for the minority class by considering the difficulty of learning for each data point in the minority class. In contrast, SMOTE generates synthetic samples by interpolating minority class instances and their nearest neighbors without considering the learning difficulty [37]. By maintaining other data processing and model construction methods consistent with the approach proposed in this study, we compare the performance of the generation algorithms.

As depicted in Figure 12, the prediction results of the ADASYN algorithm and the original SMOTE algorithm are similar, with advantages for different base classifiers. In the

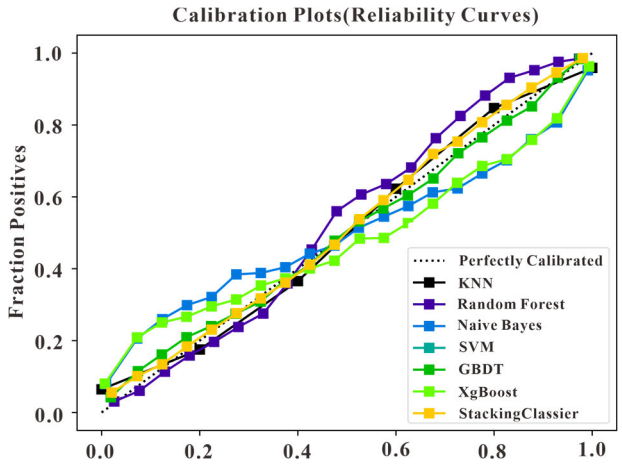

FIGURE 9. Calibration curves for different models.

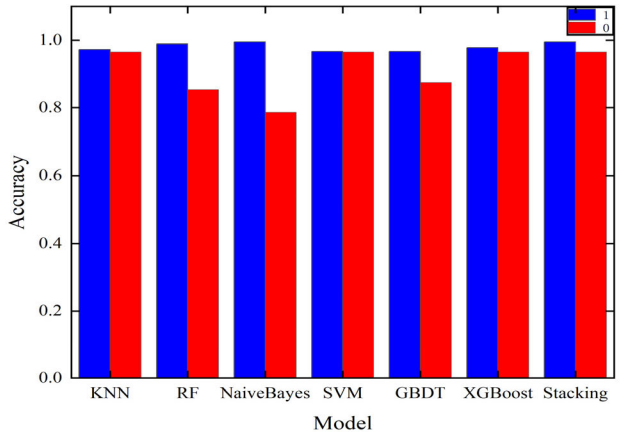

FIGURE 10. Comparison of diagnostic accuracy of different models based on real imbalanced training sets.

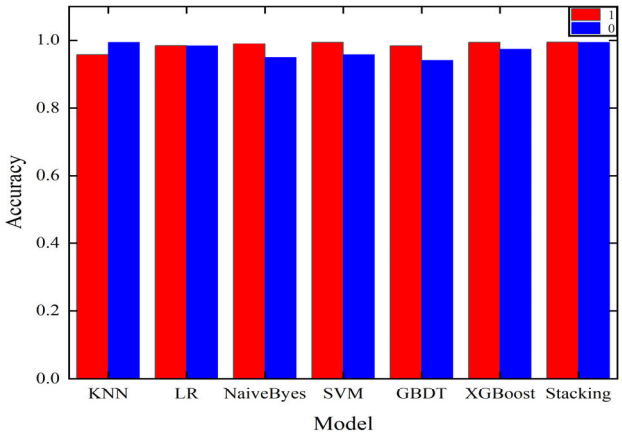

FIGURE 11. Comparison of diagnostic accuracy of different models based on synthetically balanced training sets.

Stacking model, the SMOTE algorithm performs slightly better than the ADASYN algorithm, which could be due to the dataset characteristics and the model's sensitivity to synthetic samples, leading to performance differences in

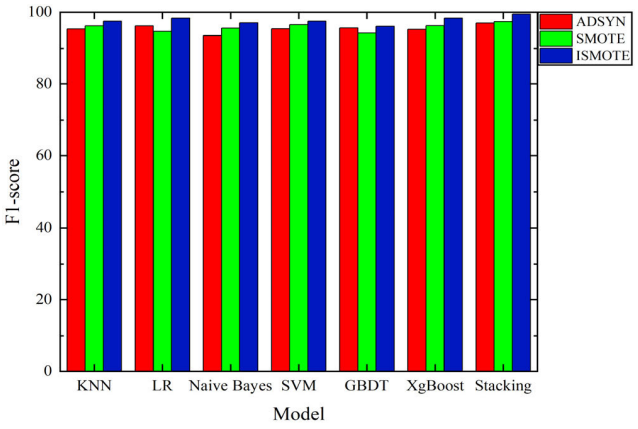

FIGURE 12. Comparison of different generation algorithms.

specific situations. However, the ISMOTE algorithm proposed in this study significantly improves prediction results, demonstrating high recognition effects for each base classification model. The proposed model achieves the highest F1-score of 99.46%, surpassing the results of all other models, indicating that ISMOTE has certain improvements compared to SMOTE and ADASYN algorithms. Additionally, the results show that employing the feature selection and feature dimensionality reduction methods mentioned in this paper can significantly enhance the model's performance.

3) THE IMPACT OF DIFFERENT DATA PROCESSING METHODS ON MODEL PERFORMANCE

To investigate the influence of different data processing methods on model performance and classification effect, we conduct experiments on the Stacking algorithm and KNN, RF, Naive Bayes, SVM, GBDT, and XGBoost under three conditions: original dataset, data processed using the ISMOTE algorithm, and data processed using ISMOTE combined with feature extraction. We calculate the accuracy, precision, recall, and F1-score of the Stacking algorithm and six basic models as evaluation indicators. The final results are shown in Figure 13.

In Figure 13, compared to the models obtained after data processing, the original dataset's classification models exhibit relatively poorer performance across all four evaluation metrics. This suggests that once subjected to feature extraction, the data possesses a certain degree of classification ability. Specifically, by employing correlation analysis and the KPCA algorithm for selecting comprehensive indicators, the correlation between features can be reduced, and some data noise can be mitigated. When adopting the same data processing method, comparing the results across various metrics reveals that the Stacking ensemble algorithm outperforms the individual base models in all metrics. This can be primarily attributed to the Stacking ensemble algorithm's ability to effectively integrate various models and maximize the performance of each algorithm. Moreover, as more efficient data processing methods are utilized, the base model

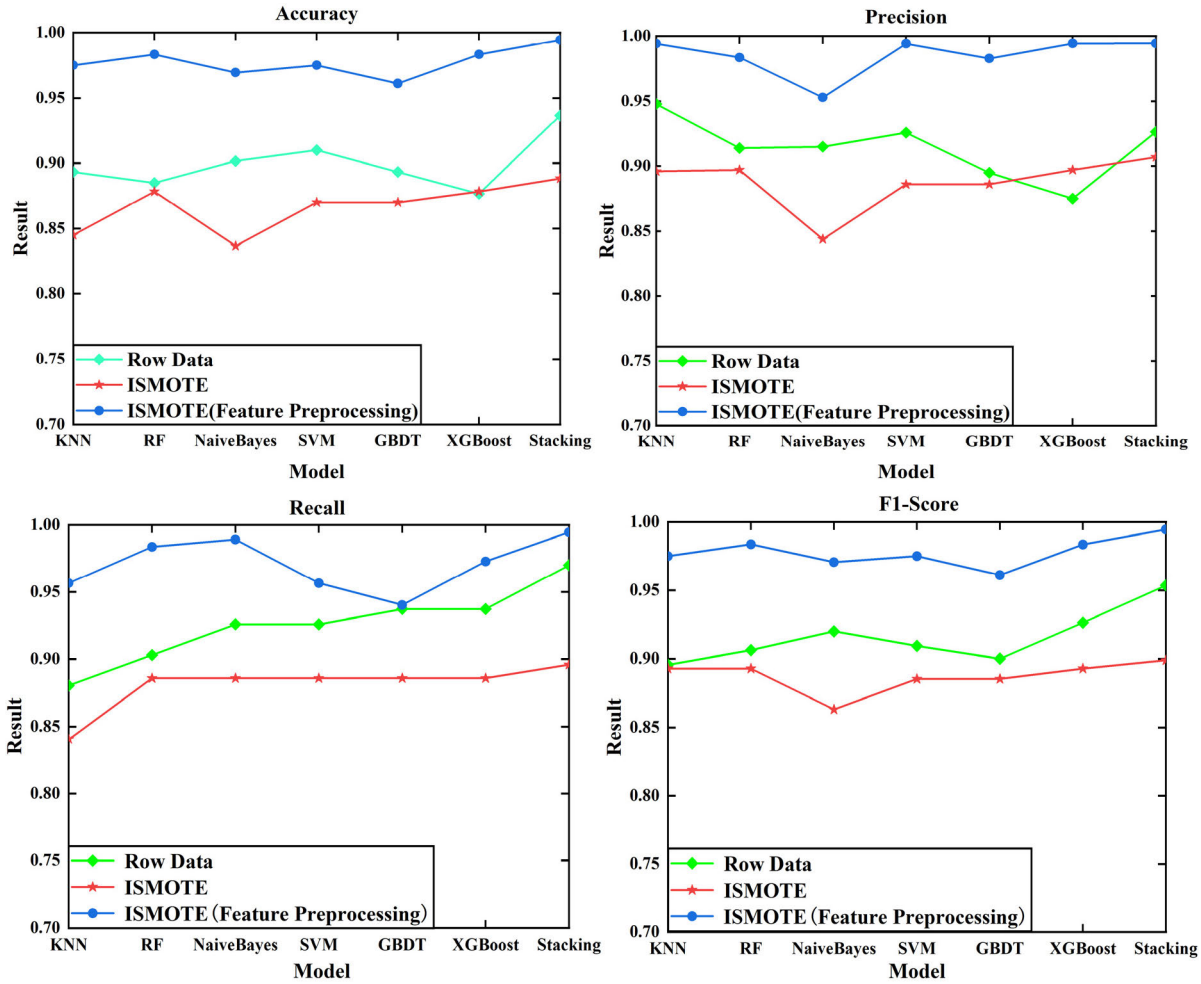

FIGURE 13. Comparison of model performance under three different processing methods.

evaluation metrics improve, further enhancing the performance of the Stacking ensemble algorithm.

In conclusion, this study demonstrates that the ISMOTE algorithm has a significant advantage in addressing imbalanced data problems. Compared to other data balancing algorithms, such as SMOTE and ADASYN, the ISMOTE algorithm exhibits better improvements in classification performance. The feature selection and feature dimensionality reduction methods mentioned in this paper also enhance model performance. Lastly, the Stacking ensemble algorithm outperforms individual base models across all evaluation metrics, indicating its strong applicability and superiority in handling imbalanced data classification tasks.

IV. DISCUSSION

Vascular vertigo/dizziness is a common neurological disorder characterized by sudden dizziness or imbalance, with its pathogenesis involving various factors, including local or systemic ischemia, hypoxia, endothelial dysfunction, and neural dysregulation. These factors may lead to vascular constriction, inflammatory response, and neuronal damage in the inner ear and vestibular nucleus, subsequently affecting vestibular function and balance perception [38]. With the continuous development of medical research, our understanding of the etiology, pathogenesis, and diagnostic methods of vascular vertigo/dizziness is constantly expanding. At present, the diagnosis of vascular vertigo/dizziness predominantly depends on the collection of medical history, physical examinations, laboratory tests (such as hemodynamics, blood lipids, blood glucose, etc.), and imaging tests (such as MRI, CT, vestibular function tests, etc.) [4].These findings necessitate further analysis and judgement from physicians. However, this process is often influenced by the physician's experience, analytical abilities, and subjective judgement, resulting in a time-consuming diagnostic process with lower accuracy. This consequently hampers swift and precise diagnosis and treatment. #Therefore, integrating intelligent diagnostic technology with existing diagnostic methods to improve the accuracy and efficiency of vascular vertigo/dizziness diagnosis has become a research focus.

Although intelligent diagnostic technology has made significant progress in various medical conditions, research

on intelligent diagnostic methods specifically for vascular vertigo/dizziness is still very limited. In this study, we significantly improved the diagnostic performance of vascular vertigo/dizziness using the ISMOTE-KPCA-STACKING algorithm. The innovative aspects of the method proposed in this study include:

1. Improved data balancing method: This study uses the ISMOTE algorithm to generate minority class data. Compared to the traditional SMOTE algorithm, the ISMOTE algorithm has adaptive boundary detection capability, allowing it to more effectively identify sparse and dense areas and focus on generating synthetic samples in sparse areas. Furthermore, the ISMOTE algorithm introduces a synthesis strategy based on K-nearest neighbors and distance weights to generate more representative samples. This data augmentation method helps improve the model's recognition of minority class samples, effectively solving the data imbalance problem.

2. Feature selection and dimensionality reduction optimization: This study uses Pearson correlation coefficients to analyze the original features' correlation, selecting significant feature indices and avoiding the impact of feature redundancy and noise on model performance. The KPCA algorithm is then introduced to perform dimensionality reduction on the selected features, mapping the feature space from high to low dimensions while retaining the data's main information. This innovative method effectively improves model performance and simplifies the calculation process, making the model more robust when dealing with complex data.

3. Ensemble learning model: This study proposes a Stacking-based ensemble learning model that integrates multiple base models and introduces a fully connected cascade neural network as a meta-layer model. The inclusion of the fully connected cascade neural network allows the ensemble algorithm to better learn the correlations between the predictions of different base models, further optimizing the overall model performance and ultimately providing the ensemble model with stronger generalization capabilities. Additionally, this study uses grid search and the Levenberg-Marquardt algorithm to optimize the parameters of the base models and meta-layer model, further improving model performance. This innovative ensemble learning model demonstrates significant advantages in vascular vertigo/dizziness diagnosis, helping to improve diagnostic accuracy and providing novel insights and effective methods for solving similar problems.

This study significantly enhances the diagnostic performance of vascular vertigo/dizziness through the use of improved data balancing methods, feature selection and dimensionality reduction optimization, and innovative ensemble learning models. These methods have advantages in vascular vertigo/dizziness diagnosis and provide new insights and effective methods for solving similar problems. To further reveal the significance of this study, we will explore the application and potential impact of this method in vascular vertigo/dizziness diagnosis and treatment from multiple perspectives.

1.Feature importance and biomarkers: Our method emphasizes identifying essential features for accurately classifying vascular vertigo/dizziness. We reveal potential biomarkers associated with vascular vertigo/dizziness by applying correlation analysis and KPCA algorithm for feature selection and dimensionality reduction. These biomarkers can help to gain a deeper understanding of the pathophysiology of vascular vertigo/dizziness and may play a crucial role in discovering new therapeutic targets.

2.Personalized medicine: Our method promotes the development of personalized medicine by effectively differentiating vascular vertigo/dizziness cases. Accurate diagnostic information allows healthcare professionals to devise personalized treatment plans based on each patient's etiology, risk profile, and clinical presentation. This will help achieve more targeted and effective treatments, reduce side effects, and improve overall patient care.

3.Machine learning interpretability: Although the method proposed in this study demonstrates good classification performance, the interpretability of machine learning models in medical settings may pose challenges. Future research should focus on integrating explainable artificial intelligence techniques into the model to enhance its transparency and interpretability. This will allow clinicians to better understand the reasoning process behind predictions, thereby increasing trust in the intelligent diagnostic process.

4. Integration with clinical practice: To successfully apply the method proposed in this study to clinical practice, seamless integration with existing workflows and decision support systems is essential. This includes developing user-friendly interfaces and tools for data input, analysis, and interpretation. Additionally, providing training and support for healthcare professionals to effectively use the intelligent diagnostic system is crucial for its successful implementation and adoption.

5.Validation in large, diverse populations: The diagnostic performance of the proposed method needs to be validated in larger, more diverse patient populations to ensure its applicability and generalizability across different healthcare settings and patient groups. This may require collaboration with multiple institutions and data sharing to achieve a sufficiently large and varied dataset for validation purposes.

6.Longitudinal studies and outcome prediction: Future research should not only focus on diagnosing vascular vertigo/dizziness but also explore the potential of the proposed method in predicting long-term outcomes and treatment responses. This can help healthcare professionals to more effectively monitor disease progression and adjust treatment strategies accordingly, further improving patient care.

In conclusion, the ISMOTE-KPCA-STACKING algorithm proposed in this study significantly improves the diagnostic performance of vascular vertigo/dizziness, offering a novel and effective approach for accurately diagnosing this complex neurological disorder. The method has the potential to contribute to personalized medicine, improve patient care, and inform future research into the pathophysiology and

treatment of vascular vertigo/dizziness. However, further research is needed to address the challenges of machine learning interpretability, integration with clinical practice, validation in diverse populations, and longitudinal outcome prediction. By addressing these challenges, we can pave the way for the broad implementation of intelligent diagnostic systems in vascular vertigo/dizziness and beyond.

V. CONCLUSION

This study proposed an efficient and accurate diagnostic model based on the ISMOTE-KPCA-STACKING algorithm to address the current challenges in diagnosing vascular vertigo/dizziness. First, we removed outliers and normalized the raw data. To address the data imbalance issue, we used the improved SMOTE algorithm to generate data, increasing the number of minority class samples to enhance the model's robustness and generalization capabilities. Next, we employed the Pearson correlation coefficient for feature correlation analysis, selecting 13 feature indicators, and introduced the KPCA algorithm for feature dimensionality reduction. KPCA demonstrates a significant advantage in class separation and effective clustering. Through the KPCA algorithm, we successfully transformed the original 13-dimensional data into 3-dimensional data, reducing signal overlap while preserving critical information, thus improving the model's performance. Finally, we constructed a Stacking ensemble algorithm model comprising various base models, including KNN, RF, Naïve Bayes, SVM, GBDT, and XGBoost, and used a fully connected cascading neural network as the meta-layer model. To optimize the parameters of the base models and the meta-layer model, we implemented grid search and the LM algorithm respectively for precise parameter adjustment. Compared to individual models, the Stacking model effectively leverages the strengths of each base model, achieving a comprehensive improvement in performance metrics such as accuracy, precision, recall, and F1-score.

The proposed ISMOTE-KPCA-STACKING algorithm demonstrates a significant advantage in diagnosing vascular vertigo/dizziness, providing an efficient and accurate solution for complex and imbalanced dataset classification tasks. The optimized data processing methods, diverse base models, and ensemble algorithm combination have broad potential in real-world clinical applications, contributing to improved patient care and treatment outcomes and offering an innovative data-driven solution in medical research. This study further validates the enormous potential of machine learning and data-driven methods in improving healthcare, optimizing diagnostic processes, and enhancing patient quality of life.

REFERENCES

[1] C. T. Zhang and J. Yi, "Cerebrovascular disease and vertigo or dizziness," *Chin. J. Stroke*, vol. 13, no. 3, p. 5, 2018, doi: 10.3969/j.issn.1673-5765.2018.03.019.

[2] M. Karatas, "Vascular vertigo: Epidemiology and clinical syndromes," *Neurologist*, vol. 17, no. 1, pp. 1–10, Jan. 2011, doi: 10.1097/NRL.0b013e3181f09742.

[3] J.-S. Kim, D. E. Newman-Toker, K. A. Kerber, K. Jahn, P. Bertholon, J. Waterston, H. Lee, A. Bisdorff, and M. Strupp, "Vascular vertigo and dizziness: Diagnostic criteria: Consensus document of the committee for the classification of vestibular disorders of the Bárány society," *J. Vestibular Res.*, vol. 32, no. 3, pp. 205–222, May 2022, doi: 10.3233/VES-210169.

[4] S.-H. Lee and J.-S. Kim, "Differential diagnosis of acute vascular vertigo," *Current Opinion Neurol.*, vol. 33, no. 1, pp. 142–149, Feb. 2020, doi: 10.1097/WCO.0000000000000776.

[5] H. L. Zhang, Y. F. Peng, D. P. Zhang, D. Li, F. X. Liu, M. Zhao, S. Yin, J. X. Liang, and T. T. Wei, "MMP-9, vertebrobasilar ectasia and vertebral artery dominance in vertigo or dizziness patients with vascular risk factors," *Frontiers Neurol.*, vol. 11, p. 931, Aug. 2020, doi: 10.3389/fneur.2020.00931.

[6] D. P. Zhang, H. R. Li, Q. K. Ma, S. Yin, Y. F. Peng, H. L. Zhang, M. Zhao, and S. L. Zhang, "Prevalence of stroke and hypoperfusion in patients with isolated vertigo and vascular risk factors," *Frontiers Neurol.*, vol. 9, p. 974, Nov. 2018, doi: 10.3389/fneur.2018.00974.

[7] D. P. Zhang, G. F. Lu, J. W. Zhang, S. L. Zhang, Q. K. Ma, and S. Yin, "Vertebral artery hypoplasia and posterior circulation infarction in patients with isolated vertigo with stroke risk factors," *J. Stroke Cerebrovascular Diseases*, vol. 26, no. 2, pp. 295–300, Feb. 2017, doi: 10.1016/j.jstrokecerebrovasdis.2016.09.020.

[8] F. Li and J. Zhuang, "Advances on clinical identification of vascular vertigo and dizziness," *China Modern Neurostatic Mag.*, vol. 23, no. 2, pp. 131–137, 2023.

[9] A. Korda, E. Zamaro, F. Wagner, M. Morrison, M. D. Caversaccio, T. C. Sauter, E. Schneider, and G. Mantokoudis, "Acute vestibular syndrome: Is skew deviation a central sign?" *J. Neurol.*, vol. 269, no. 3, pp. 1396–1403, Mar. 2022.

[10] B. B. Navi, H. Kamel, M. P. Shah, A. W. Grossman, C. Wong, S. N. Poisson, W. D. Whetstone, S. A. Josephson, S. C. Johnston, and A. S. Kim, "Application of the ABCD^2 score to identify cerebrovascular causes of dizziness in the emergency department," *Stroke*, vol. 43, no. 6, pp. 1484–1489, 2012.

[11] S. Vanni, R. Pecci, J. A. Edlow, P. Nazerian, R. Santimone, G. Pepe, M. Moretti, A. Pavellini, C. Caviglioli, C. Casula, S. Bigiarini, P. Vannucchi, and S. Grifoni, "Differential diagnosis of vertigo in the emergency department: A prospective validation study of the STANDING algorithm," *Frontiers Neurol.*, vol. 8, p. 590, Nov. 2017.

[12] A. S. Saber Tehrani, J. C. Kattah, G. Mantokoudis, J. H. Pula, D. Nair, A. Blitz, S. Ying, D. F. Hanley, D. S. Zee, and D. E. Newman-Toker, "Small strokes causing severe vertigo: Frequency of false-negative MRIs and nonlacunar mechanisms," *Neurology*, vol. 83, no. 2, pp. 169–173, Jul. 2014.

[13] M. van Smeden, J. B. Reitsma, R. D. Riley, G. S. Collins, and K. G. Moons, "Clinical prediction models: Diagnosis versus prognosis," *J. Clin. Epidemiol.*, vol. 132, pp. 142–145, Apr. 2021, doi: 10.1016/j.jclinepi.2021.01.009.

[14] G. Ahmed, M. J. Er, M. M. S. Fareed, S. Zikria, S. Mahmood, J. He, M. Asad, S. F. Jilani, and M. Aslam, "DAD-Net: Classification of Alzheimer's disease using ADASYN oversampling technique and optimized neural network," *Molecules*, vol. 27, no. 20, p. 7085, Oct. 2022, doi: 10.3390/molecules27207085.

[15] X. Yang, L. Hou, and D. Yang, "Predictors of kidney transplant rejection based on SMOTE and RNN," *Comput. Modernization*, vol. 315, vol. 11, pp. 7–11, 2021.

[16] O. Chávez-Bosquez, M. Torres-Vásquez, J. Hernández-Torruco, and B. Hernández-Ocaña, "Impacto de los algoritmos de sobremuestreo en la clasificación de subtipos principales del síndrome de guillain-barré," *Ingenius*, no. 25, pp. 20–31, Dec. 2020, doi: 10.17163/ings.n25.2021.02.

[17] L. Han, T. Yang, X. Pu, and Q. Huang, "Fuzzy logic feature selection and heterogeneous ensemble learning method for Alzheimer's disease classification," *J. Electron. Inf. Technol.*, vol. 43, no. 12, pp. 3319–3326, 2021.

[18] J. Zhang, K. Zhang, D. Lin, and Y. Chen, "A feature selection method based on discrepancy and correlation improvement," *J. Neijiang Normal Univ.*, vol. 34, no. 10, pp. 46–50, 2019, doi: 10.13603/j.cnki.51-1621/z.2019.10.009.

[19] X. He, "Quantization and application of transfer learning algorithm," Ph.D. thesis, Univ. Electron. Sci. Technol. China, China, 2021.

[20] M. Li, H. Wang, H. Long, J. Xiang, B. Wang, J. Xu, and J. Yang, "Community detection and visualization in complex network by the Density-Canopy-Kmeans algorithm and MDS embedding," *IEEE Access*, vol. 7, pp. 120616–120625, 2019, doi: 10.1109/ACCESS.2019.2936248.

[21] M. Xu, Z. Lin, and Y. Gu, "Identification of psoriasis based on PCA-SVM model," *J. Hangzhou Dianzi Univ., Natural Sci. Ed.*, vol. 38, no. 6, pp. 35–40, 2018, doi: 10.13954/j.cnki.hdu.2018.06.007.

[22] L. Ali, C. Zhu, M. Zhou, and Y. Liu, "Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection," *Exp. Syst. Appl.*, vol. 137, pp. 22–28, Dec. 2019, doi: 10.1016/j.eswa.2019.06.052.

[23] W. Su, Z. Zhang, Y. Zheng, L. Tang, and Y. Song, "Study on coronary heart disease risk prediction model based on ensemble learning," *Intell. Comput. Appl.*, vol. 12, no. 7, pp. 8–13 and 19, 2022.

[24] B. J. Kim, S.-K. Jang, Y.-H. Kim, E.-J. Lee, J. Y. Chang, S. U. Kwon, J. S. Kim, and D.-W. Kang, "Diagnosis of acute central dizziness with simple clinical information using machine learning," *Frontiers Neurol.*, vol. 12, Jul. 2021, Art. no. 691057, doi: 10.3389/fneur.2021.691057.

[25] T. Kamogashira, C. Fujimoto, M. Kinoshita, Y. Kikkawa, T. Yamasoba, and S. Iwasaki, "Prediction of vestibular dysfunction by applying machine learning algorithms to postural instability," *Frontiers Neurol.*, vol. 11, p. 7, Feb. 2020, doi: 10.3389/fneur.2020.00007.

[26] Z. Zhang, H. Liu, D. Chen, J. Zhang, H. Li, M. Shen, Y. Pu, Z. Zhang, J. Zhao, and J. Hu, "SMOTE-based method for balanced spectral non-destructive detection of moldy apple core," *Food Control*, vol. 141, Nov. 2022, Art. no. 109100, doi: 10.1016/j.foodcont.2022.109100.

[27] A. Benba, A. Jilbab, and A. Hammouch, "Voice assessments for detecting patients with Parkinson's diseases using PCA and NPCA," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 743–754, Dec. 2016, doi: 10.1007/s10772-016-9367-z.

[28] T. Yi, Y. Xie, H. Zhang, and X. Kong, "Insulation fault diagnosis of disconnecting switches based on wavelet packet transform and PCA-IPSO-SVM of electric fields," *IEEE Access*, vol. 8, pp. 176676–176690, 2020, doi: 10.1109/ACCESS.2020.3026932.

[29] F.-S. Chen, H.-Y. Jiang, and Z. Jiang, "Prediction of drug-pathway interaction pairs with a disease-combined LSA-PU-KNN method," *Mol. BioSyst.*, vol. 13, no. 12, pp. 2583–2591, 2017, doi: 10.1039/C7MB00441A.

[30] M. Song, H. Jung, S. Lee, D. Kim, and M. Ahn, "Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm," *Brain Sci.*, vol. 11, no. 4, p. 453, Apr. 2021, doi: 10.3390/brainsci11040453.

[31] S. Narayan and E. Sathiyamoorthy, "Early prediction of heart diseases using naive Bayes classification algorithm and Laplace smoothing technique," *Int. J. Grid High Perform. Comput.*, vol. 14, no. 1, pp. 1–14, Jan. 2023, doi: 10.4018/IJGHPC.316157.

[32] A. Abdulkadir, B. Mortamet, P. Vemuri, C. R. Jack, G. Krueger, and S. Klöppel, "Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier," *NeuroImage*, vol. 58, no. 3, pp. 785–792, Oct. 2011, doi: 10.1016/j.neuroimage.2011.06.029.

[33] J. Zhang, D. Xu, K. Hao, Y. Zhang, W. Chen, J. Liu, R. Gao, C. Wu, and Y. De Marinis, "FS-GBDT: Identification multicancer-risk module via a feature selection algorithm by integrating Fisher score and GBDT," *Briefings Bioinf.*, vol. 22, no. 3, May 2021, Art. no. bbaa189, doi: 10.1093/bib/bbaa189.

[34] A. Ogunleye and Q. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: 10.1109/TCBB.2019.2911071.

[35] G. Deshpande, P. Wang, D. Rangaprakash, and B. Wilamowski, "Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2668–2679, Dec. 2015, doi: 10.1109/TCYB.2014.2379621.

[36] N. Narayanan, G. Beyene, R. D. Chauhan, M. A. Grusak, and N. J. Taylor, "Stacking disease resistance and mineral biofortification in cassava varieties to enhance yields and consumer health," *Plant Biotechnol. J.*, vol. 19, no. 4, pp. 844–854, Apr. 2021, doi: 10.1111/pbi.13511.

[37] J. Beinecke and D. Heider, "Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making," *BioData Mining*, vol. 14, no. 1, p. 49, Dec. 2021, doi: 10.1186/s13040-021-00283-6.

[38] T. Inui, T. Kuriyama, S.-I. Haginomori, K. Moriyama, T. Shirai, Y. Ayani, Y. Inaka, M. Araki, and R. Kawata, "Different results of vestibular examinations and blood flow in cases with transient vascular vertigo/dizziness with or without central nervous system symptoms," *Acta Oto-Laryngologica*, vol. 142, nos. 9–12, pp. 685–690, Dec. 2022, doi: 10.1080/00016489.2022.2134587.

DENGQIN SONG was born in Bijie, Guizhou, China, in 1995. She received the bachelor's degree from the Guizhou University of Traditional Chinese Medicine, in 2019. She is currently pursuing the master's degree with the Hubei University of Chinese Medicine. Her research interests include the prevention and treatment of geriatric diseases and digestive system disorders through traditional Chinese medicine and medical artificial intelligence. She has demonstrated solid professional knowledge and research potential in these fields and looks forward to making significant breakthroughs in her future academic career.

TONGQIANG YI was born in Jining, Shandong, China, in 1992. He received the master's degree in software engineering from Xi'an Jiaotong University. He is currently pursuing the Ph.D. degree with the School of Power and Mechanical Engineering, Wuhan University. His primary research interests include pattern recognition, intelligent fault diagnosis, and the innovative application of artificial intelligence in interdisciplinary fields.

QINGWEI XIANG was born in Hubei, China, in 1983. He received the Graduate degree from the Hubei University of Chinese Medicine under the Integrative Medicine Program, and the master's degree. He is currently an associate chief physician. He is also with the Department of Geriatrics, Hubei Provincial Hospital of Traditional Chinese Medicine. He has more than ten years of experience in clinical practice, research, and teaching. He has published over ten papers in core, national, and provincial journals. He has led or participated in several national and provincial-level research projects. He also serves as a Committee Member of the General Practitioner Professional Committee of Hubei Provincial Association of Chinese Medicine, a member of the Osteoporosis Professional Committee of Wuhan Geriatrics Society, and a member of the Neuroimmunomyopathy Group of Wuhan Medical Association's Neurology Branch.

HONGCI CHEN was born in Hubei, China, in 1974. She received the master's degree from the Hubei University of Traditional Chinese Medicine. She is currently a chief physician and a master's degree supervisor. She serves as a Standing Committee Member of the First Committee of the Health and Health Education Professional Committee of Hubei Association of Geriatric Healthcare, a Standing Committee Member of the Hubei Alliance of the Geriatric Comprehensive Assessment Collaborative Innovation Alliance, a Standing Committee Member of the Second Committee of Spleen and Stomach Diseases Professional Committee of Wuhan Traditional Chinese Medicine Association, a Committee Member of the First Committee of Geriatric Medicine Professional Committee of the Chinese Female Physicians Association, and the Director of the Fourth Executive Council of Hubei Medicinal Diet and Food Therapy Research Association. She has extensive professional knowledge and experience in the fields of traditional Chinese medicine and geriatric medicine, making significant contributions to the medical community.