applied sciences

MDPI

Article

Ischemic Stroke Prediction by Exploring Sleep Related Features

Jia Xie [1,2], Zhu Wang [1,*], Zhiwen Yu [1,*], Bin Guo [1] and Xingshe Zhou [1]

[1] School of Computer Science, Northwestern Polytechnical University, 1 Dongxiang Road, Chang'an District, Xi'an 710021, China; xiejia@mail.nwpu.edu.cn (J.X.); guob@nwpu.edu.cn (B.G.); zhouxs@nwpu.edu.cn (X.Z.)
[2] School of Electronic Information and Artificial Itelligence, Shaanxi University of Science and Technology, Xi'an 710021, China
* Correspondence: wangzhu@nwpu.edu.cn (Z.W.); zhiwenyu@nwpu.edu.cn (Z.Y.); Tel.: +86-29-8843-1531 (Z.W.)

Abstract: Ischemic stroke is one of the typical chronic diseases caused by the degeneration of the neural system, which usually leads to great damages to human beings and reduces life quality significantly. Thereby, it is crucial to extract useful predictors from physiological signals, and further diagnose or predict ischemic stroke when there are no apparent symptoms. Specifically, in this study, we put forward a novel prediction method by exploring sleep related features. First, to characterize the pattern of ischemic stroke accurately, we extract a set of effective features from several aspects, including clinical features, fine-grained sleep structure-related features and electroencephalogram-related features. Second, a two-step prediction model is designed, which combines commonly used classifiers and a data filter model together to optimize the prediction result. We evaluate the framework using a real polysomnogram dataset that contains 20 stroke patients and 159 healthy individuals. Experimental results demonstrate that the proposed model can predict stroke events effectively, and the Precision, Recall,Precision Recall Curve and Area Under the Curve are 63%, 85%, 0.773 and 0.919, respectively.

Keywords: ischemic stroke; sleep EEG; sleep cycle; sleep stage

Citation: Xie, J.; Wang, Z.; Yu, Z.; Guo, B.; Zhou, X. Ischemic Stroke Prediction by Exploring Sleep Related Features. *Appl. Sci.* **2021**, *11*, 2083. https://doi.org/10.3390/app11052083

Academic Editor: José Ignacio Serrano

Received: 20 December 2020
Accepted: 20 February 2021
Published: 26 February 2021

1. Introduction

Ischemic stroke is a medical condition, which occurs once the arteries to the brain become blocked, resulting in cell death. [1]. Each year, in the United States, approximately 795,000 people suffer from a stroke, and about 600,000 of these are first attacks, 185,000 are recurrent attacks [2]. In medicine, stroke has two subspecies: ischemic stroke and hemorrhagic stroke. Compared with hemorrhage, ischemic stroke accounts for about 87% of all strokes [3]. It occurs when a blood vessel is blocked by a blood clot which means the blood cannot reach the brain. Therefore, the cerebral blood flow (CBF) of patients suffering from the ischemic stroke is lower than those of health [4]. Obviously, ischemic stroke damages brain cells and hence leads to the degeneration of the motor function accompanied with depression and anxiety [5]. If ischemic stroke can be predicted in a timely fashion, the quality of life and the health level of the public can be improved significantly.

To predict stroke accurately, two key problems should be solved. First, when there are no apparent symptoms, can we extract effective ischemic stroke predictors to diagnose or predict stroke? Second, for ease of use, can we extract efficient stroke predictors from data that can be obtained during people's daily lives in a non-intrusive manner?

By now, researchers have conducted numerous studies to extract useful ischemic stroke predictors, which can be roughly grouped into three categories. The first category of studies focuses on extracting features manually based on expert medical knowledge. For example, Sedghi et al. [6] and Lumley et al. [7] constructed a stroke prediction model by manually selecting 16 features. Although the effectiveness of some clinical features [8–10]

had been verified by existing studies, how to extract efficient features from continuous physiological signals is still an open issue. The second category of studies aims to extract features from Electrocardiograph (ECG), Computerized Tomography Scanner (CT-Scan), or Magnetic Resonance Imaging (MRI) investigation. Particularly, based on ECG, CT-Scan and MRI, the detailed structures of soft tissues, such as brain and heart, can be observed more clearly. For example, Sajjadi et al. [11] and other authors [12–14] verified the usefulness of the CT-Scan and MRI data of the brain in early detection of acute ischemic stroke. While existing approaches have achieved a satisfactory stoke detection performance, they can only detect stroke that will occur soon or has already occurred. To enable early prediction of stroke occurrence, we need more efficient approaches. The third category of studies extracts features utilizing machine learning techniques, which were widely used to investigate the relationship between the symptoms and stroke incidents. For instance, Khosla et al. [15,16] and Kasabov et al. [17] proposed an integrated machine learning method to predict stroke based on different environmental factors. All these studies mainly focused on detecting stroke incidents by mining features from discrete data, rather than leveraging continuous data that contain much more useful information.

In summary, existing stroke prediction methods have not fully solved the problem of early stroke prediction, and the main shortcoming is that they did not fully exploit unobtrusive sensing techniques to collect long-term health-related time series data during daily life (such as continuous health data that can be collected during sleep), which may leave out useful hidden information when predicting stroke. Specifically, in this paper, we aim to address the early prediction problem of ischemic stroke by exploring sleep-related time series data. On one hand, ischemic stroke occurs most often shortly after awakening in the morning hours [18], and a number of studies have reported a sleep circadian variation in the occurrence of stroke [19,20]. For example, patients' sleep was found to be markedly altered compared with a normal group, specific performance in a higher amount of slow wave sleep was recorded, whereas fast wave sleep was found to be deeply suppressed (such as stage 0 being increased, however stage 4 being reduced) [21,22], coupled to present electroencephalograph (EEG) abnormalities such as an increase in delta frequency, sometimes an increase and at other times a decrease in theta activity and/or a decrease in alpha and beta frequencies [23]. Ma et al. [24] had found that participants with rapid eye movement (REM) sleep behavior disorder were approximately 1.5 times more likely to develop stroke, including the ischemic type. In another similar study, it was also proven that sleep disorders were highly prevalent in patients at risk for stroke and may be modifiable risk factors for stroke [22]. Finnigan et al. [25] certified that the delta/alpha power ratio (DAR) and the (delta + theta)/(alpha + beta) ratio (DTABR) exhibited optimal classifier accuracy between ischemic stroke patients and control groups. On the other hand, EEG offers a continuous, real-time, non-invasive measure of brain function [26] and it is the most sensitive neuro-diagnostic tool for detecting cerebral ischemia and can show changes of ischemic stroke within minutes of onset [27]. EEG abnormalities are closely tied to CBF [27]. When normal CBF of 50 mL/100 g/min declines to 25 mL/100 g/min, EEG will first lose faster frequencies ($\geq$8 Hz), and then if CBF continues to drop to 18 mL/100 g/min, which is a crucial ischemic threshold, the slower frequencies (4 Hz to 7 Hz) of EEG will increase. Figure 1 shows the relationship of EEG and ischemic stroke.

Therefore, by extracting sleep-related features (including sleep structure and sleep EEG) to predict ischemia, the first key problem can be solved. In particular, these sleep features do not compete with elegant clinical trials, but the sleep information it adds is complementary to other information. In addition, wearable devices (like smart waist-belts), can be utilized to collect daily sleep-related data, which are as accurate as clinical equipment. Furthermore, EEG is a well-known, user-friendly examination and is less complex than the imaging modality [28]. It is usually recorded by using 2 to 19 electrodes, the numbers of which are set according to the needs of the patient, placed upon the scalp according to the International 10–20 System. Since the recruitment of stroke participants is extremely expensive and time-consuming, this study will not collect experimental datasets by using

portable data acquisition devices. Nevertheless, it is notable that we can collect datasets which are similar to the one utilized in this study through portable data acquisition devices, and the proposed method can be directly used for data analysis and stroke prediction. As for the second problem, we will solve it in the the rest of the paper.

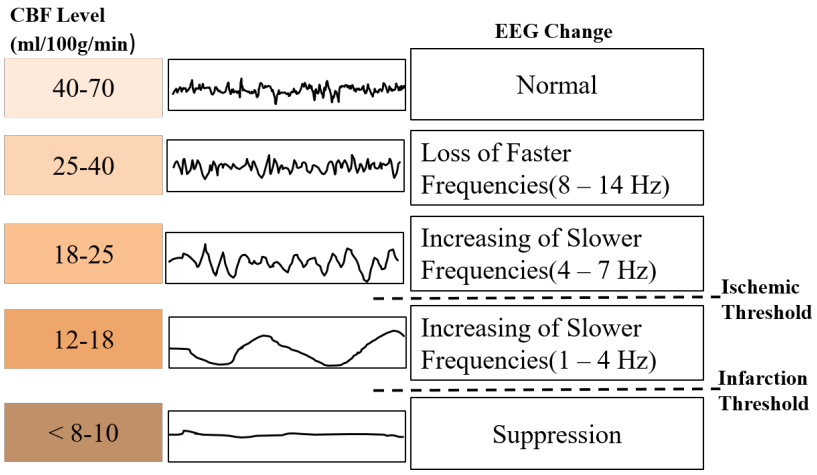

Figure 1. The relationship of electroencephalograph (EEG) change and Ischemic stroke [26].

There has been little work that has explored sleep-related features and further combined them with other clinical features (e.g., high blood pressure, cholesterol [7]) to predict the occurrence of stroke. In our work, we present a novel ischemic stroke prediction approach, which integrates clinical features with fine-grained sleep-related features for stroke risk prediction. Specifically, to fully characterize the sleep pattern of a high-risk population of ischemic stroke, we extracted both sleep structure related and sleep EEG related features. To sum up, the contributions of this study are three-fold:

First, compared with most of the previous studies, which merely used sleep data from a few minutes to tens of minutes, we had leveraged all night's sleep data to make this prospective study. Specifically, we extracted both sleep structure related and sleep EEG related features to characterize the ischemic stroke pattern more comprehensively.

Second, since the consequences of stroke incidents are quite severe, we should try our best to correctly identify people who will have a stroke in the short run. In other words, we should keep the false negative rate of the proposed model as low as possible. Meanwhile, compared with the false negative rate, a slightly higher false positive rate will not result in severe loss. Therefore, we developed a novel stroke risk prediction model, which consists of two steps. The first one is a machine learning model to obtain basic prediction results, and the second one is a pre-selection model to further optimize the false negative rate of stroke prediction.

Third, the proposed stroke prediction model is evaluated by a using 10-fold cross-validation technique based on a real polysomnogram (PSG) dataset. Experimental results demonstrate that the proposed model (i.e., Support Vector Machine (SVM)+pre-selection) outperforms the baseline method (i.e., the model based on basic SVM) by 15% in terms of the true positive rate (TPR).

The rest of the paper is organized as follows. Section 2 reviews the related work, followed by problem formulation and approach overview in Section 3. In Section 4, we describe the process of feature extraction. Section 5 details the elaboration of the proposed prediction model. We analyze the experimental results in Section 6, followed by conclusion and possible future works in Section 7.

2. Related Work

2.1. Ischemic Stroke Predictor

Clinical predictors may be useful in clinical practice to support ischemic stroke treatment. For example, Singer et al. [29] used the original Anticoagulation and Risk Factors in an Atrial Fibrillation cohort to predict ischemic stroke and showed improvement in predicting severe events. In another similar article [30], authors found that CHADS2 could quantify the risk of ischemic stroke for patients who had Atrial Fibrillation. On the other hand, through studying 11 hemostatic markers, Ann Smith et al. [31] discovered that plasminogen activator inhibitor-1, Factor VII coagulant activity, D dimer and Fibrinogen had potential to increase the prediction of ischemic stroke in middle-aged men. In addition, the stroke risk score-based approach has also been proposed. For instance, the best-known stroke risk score is the Framingham Stroke Risk, which was developed as a part of the Framingham Heart Study and used to estimate 10-year cardiovascular risk [32]. Besides these clinical characteristics, researchers also assessed the predictivity of ischemic stroke in a community study. For example, based on four US communities' participants, Chambles et al. [33] found that several nontraditional factors (e.g., waist:hip ratio, high density lipoprotein cholesterol, alcohol consumption and so on) can significantly improve the performance of ischemic stroke prediction over a risk score that only included traditional factors (e.g., current smoking status, diabetes mellitus, systolic blood pressure and so on). However, these approaches, which rely on manually selected features, are not suitable for obtaining patterns from huge datasets and would result in poor identification performance. A better choice is to adopt machine learning-based approaches, which can discover hidden stroke patterns from enormous datasets in a relatively cost-effective manner.

2.2. Prediction Models

With the rise of new techniques and the ever-increasing medical data available, the problem of stroke risk prediction has been studied from different aspects. For example, Chien et al. [34] and Jee et al. [35] used a regression model (i.e., the Cox proportional hazards model) to predict the risk of stroke and identify individuals at high risk of stroke. Similarly, Khosla et al. [15] used SVM to predict the risk of stroke within 5 years using the CHS dataset, and obtained a better result compared with the Cox model. For predicting ischemic stroke, Arslan et al. [36] assessed different medical data mining approaches and found that SVM produced the best predictive performance compared with the two other models. More recently, deep neural networks had been investigated for creating predictions from electronic health records. For instance, Goyal et al. [37] employed a Recurrent Neural Network architecture with Long Short-Term Memory hidden units for the prediction of stroke. While existing machine learning models can achieve effective prediction, most of the used clinical features that cannot be obtained conveniently and continuously. In addition, none of them had tried to control the false negative rate of stroke prediction. To tackle these problems, we first extracted a set of features from sleep data which can be collected conveniently during people's daily lives. Afterwards, we adopted a two-step prediction model to make the false negative rate as low as possible. Compared with our previous work [38], which exploited sleep structure data to extract fine-grained sleep cycles and stages features; in this paper, in addition to sleep structure, we also utilized a sleep EEG stream to extract sleep EEG related features (included relative power and three nonlinear features according to different sleep cycles and stages) and produced good results.

2.3. EEG and Ischemic Stroke

Due to the physiologic coupling of EEG with cerebral blood flow [39], ischemic stroke may result in EEG rhythm changes, such as polymorphic delta, attenuation of fast activity, sleep spindles, and so on [40]. Therefore, EEG adds value to early diagnosis, outcome prediction, patient selection for treatment, clinical management, and seizure detection in acute ischemic stroke [27]. Compared with EEG, serial neurological exams and

imaging (such as computed tomography (CT) scans, which do not reveal early ischemic stroke [41], and Magnetic Resonance Image (MRI), which is expensive and often has a lag time of hours before detecting ischemic stroke [42]) are only capable of detecting irreversible cellular damage [40,43]. By contrast, EEG is widely available, inexpensive, and can show changes of acute ischemic stroke within minutes of onset [27]. In addition to early detection of ischemic stroke, EEG can also be helpful during the diagnosis and evaluation. Molnár et al. [44] found that with increasing absolute delta, theta, and Omega-complexity in these frequency bands, higher theta/beta ratios and decreased relative beta activity were found in the side of the infarct. Daroff et al. [45] pointed out that although there were a few theta waves among healthy persons, theta and delta waves were mostly found in the pathological state of the brain, and persistent pleomorphic slow activity is closely related to local cerebral lesions (e.g., infarction, hemorrhage, and tumor). Therefore, continuous monitoring of ischemia-related changes in EEG would be of great value to neurologists in guiding therapy [46,47]. Therefore, on one hand, we used whole night sleep EEG to determine the sleep morphology and its clinical predictive value. On the other hand, we divided each sleep EEG into different phases (according to different sleep cycles and sleep stages) to characterize ischemic stroke pattern from another perspective.

3. Problem Statement and Approach Overview

3.1. Problem Statement

PSG provides a data-rich source for understanding and measuring sleep brain activities [48]. For instance, based on the well-known Rechtschaffen and Kales scoring criteria [49], people can leverage these neurophysiological signals for standard sleep staging. The main stages are wakefulness, rapid eye movement (REM) sleep and non-rapid eye movement (NREM) sleep, where NREM is further divided into four stages from the lightest sleep stages 1 and 2 to the deepest sleep stages 3 and 4. Specifically, sleep stage 1 is characterized by theta waves (the amplitude is 50 to 100 micro volts); the characteristics of stage 2 sleep are termed sleep spindles (the amplitude is 50 to 150 micro volts); stage 3 is considered spindle waves and slow waves sleep (the amplitude is 100 to 150 micro volts); stage 4 has the same attributes as stage 3, but more than 50% of the waves are slow waves and delta waves (the amplitude is 100 to 200 micro volts).

In this paper, we leverage features extracted from psychological sleep data for ischemic stroke prediction. The formulation below represents the time series of sleep data:

$$SS = \{sc_1, sc_2, \ldots, sc_n\}, \tag{1}$$

where n is the number of sleep cycles. Specifically, each sleep cycle sc_i ($i \in [1, n]$) can be denoted as:

$$sc_i = (s0_i, s1_i, s2_i, s3_i, s4_i, s5_i), \tag{2}$$

where $s0_i$ is the awake stage of the sleep cycle sc_i, $s1_i$ and $s2_i$ represent the shallow sleeping stages, $s3_i$ and $s4_i$ denote the stage 3 and 4 of deep sleeping, and REM is represented by $s5_i$.

The average time span of a sleep cycle lasts about 90 min [50], and most people have four or five sleep cycles overnight with each starting with NREM sleep (i.e., $s1$, $s2$, $s3$, $s4$) and then followed by a short REM sleep (i.e., $s5$). As the sleeping process goes on, the duration of stage 3 and stage 4 decreases while the duration of REM increases. As a result, stage 3 to 4 takes up a larger proportion in the earlier part of the night, and REM sleep occupies a greater part of the later part of the night [51]. Figure 2 shows the typical structure of a whole night's sleep stages and cycles.

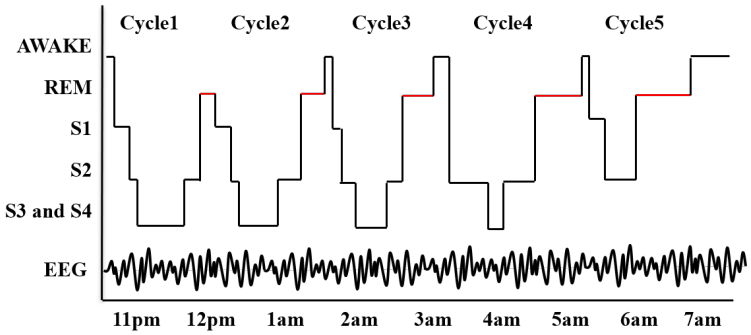

Figure 2. A representative of sleep cycles and concomitant EEG signal.

A number of studies have proven that more than one fourth of stroke occurs during sleep accompanied with a circadian variation [52]. In addition, the frequency pattern of EEG was very sensitive to changes in neuronal function resulting from ischemia [53]. As a result, it is reasonable to assume that the fusion of sleep stage and EEG change related features can help to characterize ischemic stroke from certain aspects, and even predict the occurrence of stroke events. In order to enable early detection of stroke, we propose to extract effective features from the sleep data series (including sleep cycles, sleep stages, and sleep EEG) to distinguish patients with stroke in early phases from healthy individuals. The problem is defined as follows.

Problem Formalization- Given a time series of sleep data stream of either a healthy individual or a patient TSS = $\{TSC_1, TSC_2, \ldots, TSC_n\}$, and $TSC_i = (TS0_i, TS1_i, TS2_i, TS3_i, TS4_i, TS5_i)$, our aim is to extract features from different dimensions and then build a model to predict ischemia.

3.2. Approach Overview

As shown in Figure 3, we first use raw data from two different sources, namely physiological data obtained by special medical equipment and sleep data gathered from portable sensing devices (e.g., a standard EEG according to the international 10 to 20 system). Then, we extract stroke predictors, including clinical statistical features, sleep structure related features, and sleep EEG related features. Afterwards, we propose a two-step stroke prediction method, in which we combined the regular machine learning algorithm with the pre-selection model. More specifically, during the stroke pattern analysis process, we obtain primary results using the machine learning algorithm at the beginning, and then apply the pre-selection model to optimize them, aiming at minimizing the false negative rate without severely affecting the false positive rate.

In the following sections, the extracted features will be first introduced and then the prediction model will be described in more detail.

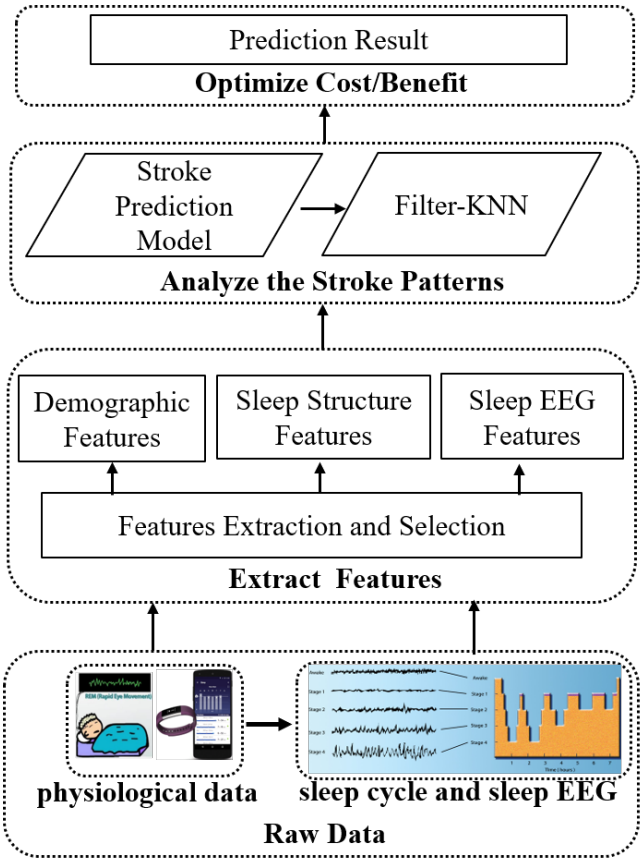

Figure 3. A general view of the prediction model.

4. Feature Extraction

In this section, we present the details of the extracted features: clinical features and polysomnogram features (including sleep structure related features and EEG sleep related features).

4.1. Clinical Features

The selection of relevant features plays a crucial role in constructing an efficient prediction model. There are numerous attributes in the Sleep Heart Health Study (SHHS) dataset, including clinical history, demographic information, physical and biomedical measurements [15]. However, only a few of them are relevant to ischemia. Existing methods usually utilize several risk factors identified by clinical studies to predict stroke. In this study, we also utilize such risk factors as clinical features, such as sex, diastolic pressure, systolic pressure, hypertension, diabetes mellitus, atrial fibrillation, race, smoking, peripheral vascular, left ventricular hypertrophy, total cholesterol, aspirin and high-density cholesterol.

4.2. Sleep Structure Related Features

As for a healthy subject, the sleep cycle usually starts with the NREM stage, and followed by the REM stage. These stages will alternate throughout the whole night periodically. Particularly, most of the slow-wave NREM stages occur in the first half of the night. For example, $s3$ and $s4$ sleep stages occupy much less time in the second cycle than in the first cycle and even disappear in later cycles. Moreover, the REM sleep stage (i.e., $s5$) may only last about 1 to 5 min in the first cycle and will gradually become longer through the night.

Sleep-related problems are very clinical among stroke patients, and even more than 50% of stroke survivors suffer from at least one certain type of sleep problem [54–56]. Specifically, sleep disorder is a typical sleep problem among stroke patients, which includes

hypersomnia, insomnia, sleep-disordered breathing, etc. All of these symptoms may lead to dyssomnias, parasomnias, circadian rhythm sleep disorders etc. [1]. While these symptoms are observed among stroke patients, we have reason to assume that these features also exist among subjects who will suffer from ischemia in a near future. Therefore, in this study, we will attempt to build a stroke prediction model by extracting and leveraging sleep related features.

To investigate the effects of sleep disorders on stroke prediction, we extract sleep structure related features from two aspects. In particular, the first aspect is sleep quality-related features, as more than one third stroke patients have reduced sleep time, insomnia [57], increased awake times [58], low sleep efficiency [21,59], shorter rapid eye movement-sleep latency and higher sleep latency [60]. The second aspect is sleep trend-related features, as it has been proven that the deep sleep stage of stroke patients will decrease, while the shallow sleep stage of stroke subjects will increase [21,59]. Therefore, similar to our previous work [38], we extract a number of sleep structure related features, including features total sleep time (TST), sleep efficiency (SE), wake after sleep onset (WASO), awake times (AT), sleep latency (SL), rapid eye movement-sleep latency, FF Trend 1, TF Trend 1 and SF Trend1.

4.3. EEG Sleep Related Features

EEG, a record of the oscillations of brain electric potential [61], is useful for the analysis of functional changes due to the regional brain pathology of ischemic stroke [28]. The EEG patterns of wakefulness and sleep usually differ from each other markedly, as well as the patterns of different levels of sleep [62]. In general, EEG, the frequencies of which range from 0.5 to 100 Hz, is decomposed into five rhythms: delta (1 to 4 Hz), theta (4 to 7 Hz), alpha (7 to 12 Hz), beta (12 to 25 Hz) and gamma (above 25 Hz) [63]. In this paper, we are only interested in the components of EEG signals below 25 Hz; therefore, the EEG signals were band-limited to the desired 1 to 25 Hz range by using wavelet transform-based methods. In the rest of this section, we present features from two aspects: frequency domain features (including relative power, brain symmetry index) and nonlinear features (including sample entropy, detrended fluctuation analysis and Lempel–Ziv complexity).

Power spectral density, which is performed using wavelet packet decomposition to calculate relative power (RP), power ratio, etc., is the most clinical feature for analyzing EEG data to characterize the changes of EEG patterns [64]. In this work, we quantify the relative power of each selected frequency band, i.e., delta, theta, alpha and beta, to distinguish the EEG of ischemic stroke patients from that of the healthy subjects. The relative power can be formally defined as follows [65].

$$RP_{(F_i,C_x)} = \frac{\sum selected\ frequency\ band\ energy}{\sum Total\ EEG\ range\ energy}. \tag{3}$$

where F_i represents the four different frequency bands of the *EEG*, i.e., i $\in \{D, T, A, B\}$; C_x is the channel of the EEG, and in this study we used two different channels (C_3 and C_4), i.e., $x \in \{3, 4\}$.

Furthermore, in order to obtain more fine-grained features, we divided all these frequency bands according to different sleep cycles and sleep stages and calculate the relative power as follows.

$$RP_{(F_i,s_j,C_x)} = \frac{E_{(F_i,s_j,C_x)}}{E_{(F_i,C_x)}}. \tag{4}$$

$$RP_{(F_i,s_j,cyc_k,C_x)} = \frac{E_{(F_i,s_j,cyc_k,Cx)}}{E_{(F_i,cyc_k,C_x)}}. \tag{5}$$

where s_j is the j-th sleep stage and $j \in \{0, 1, 2, 3, 5\}$; cyck is the k-th sleep cycle and $k \in \{1, 2, 3, 4, 5\}$. $E_{(F_i,s_j,C_x)}$ is the energy of frequency band F_i on the EEG channel C_x during the j-th sleep stage; $E_{(F_i,C_x)}$ is the total energy of frequency band F_i on the EEG channel C_x;

$E_{(F_i,s_j,cyc_k,C_x)}$ is the energy of frequency band F_i on the EEG channel C_x during the j-th sleep stage of the k-th sleep cycle; $E_{(F_i,cyc_k,C_x)}$ is the total energy of frequency band F_i on the EEG channel C_x during the k-th sleep cycle.

Subsequently, the power ratio of delta/alpha (DAR), theta/beta (TBR), the (delta + theta)/(alpha + beta) power ratio (DTABR), and the (theta − delta)/(alpha − beta) power ratio (TDABR) were also calculated, as the usefulness of these features had been proven by existing studies. For example, Claassen et al. [40] proved that the DAR demonstrated the strongest association with cerebral ischemia. Finnigan et al. [44] pointed out that DTABR was obviously linked to DAR, which may sometimes be informative to ischemia monitoring. In addition, higher TBR was found in the side of the infarct [44], and power ratios such as TDABR may be useful for detecting early and subtle ischemic EEG changes [27]. All these features are also divided according to different sleep cycles and sleep stages in order to detect the changes consequent to the stroke. These features include: $DAR_{(F_i,s_j,cyc_k,C_x)}$, $TBR_{(F_i,s_j,cyc_k,C_x)}$, $DTABR_{(F_i,s_j,cyc_k,C_x)}$ and $TDABR_{(F_i,s_j,cyc_k,C_x)}$.

The Brain Symmetry Index (BSI): BSI, which calculates the brain symmetry, is one of the popular EEG-derived parameters used in the research field for the purposes of stroke prognostication [66]. It is defined as the mean of the absolute value of the difference in mean hemispheric power in the frequency range from 1 to 25 Hz [67].

$$BSI = \frac{1}{NM}\sum_{j=1}^{M}\left|\sum_{i=1}^{N}\frac{R_{ij}-L_{ij}}{R_{ij}+Lij}\right|. \quad (6)$$

where R_{ij} and L_{ij} are the power spectral density obtained using Welch's method of the right and the left hemisphere, respectively. While M and N are the total number of Fourier coefficients j = 1, 2,. . . , M and total number of electrode pairs i = 1,2,. . . , N [64]. Specifically, BSI is also divided according to different sleep stages and sleep cycles, i.e., we obtain a set of features as $BSI_{(s_j,cyc_k,C_x)}$.

In addition to the existing researches concentrated on power spectral density, which has been revealed to be an efficient indicator, it is well accepted that the brain is a complex system and nonlinear measures should be taken into account for modeling and analysis [68]. Therefore, we also extracted three nonlinear features, including Sample Entropy, Detrended fluctuation analysis and Lempel–Ziv complexity.

Sample Entropy (SampEn): SampEn is an effective metric to improve the approximate entropy method [69], and it characterizes the complexity and regularity of short-time series and has been widely used in bioinformatics [70]. Moreover, SampEn is not sensitive to the noise, making it appropriate for analyzing the EEG data [71]. Specifically, there are two important parameters when calculating SampEn, which are m (the embedding dimension) and r (the tolerance threshold). In this work, we set m = 2 and r = 0.2 * standard deviation, according to the empirical results.

Detrended fluctuation analysis (DFA): DFA, which eliminates the trends in time series, is a method for analyzing variability of biomedical signals [72], and its result indicates the EEG fluctuation by scaling exponent [73]. In this work, we use DFA to reveal the long-term inner correlations of the EEG data by setting the window size as 16 to 512.

Lempel–Ziv complexity (LZC): LZC is calculated based on the method introduced by Lempel and Ziv [74]. It characterizes the disorder of a time series by testing the emergence rate of a new model of the biological signals [75], especially the EEG data [76]. In this paper, EEG data are segmented into consecutive sequences of 5 seconds and a binarization process is conducted according to the mean-value of the EEG time series.

The same as the power spectral density, we calculate all these nonlinear features according to different sleep cycles and sleep stages, including $\mathrm{SampEn}_{(s_j,cyc_k,C_x)}$, $\mathrm{SampEn}_{(F_i,s_j,cyc_k,C_x)}$, $\mathrm{DFA}_{(s_j,cyc_k,C_x)}$, $\mathrm{LZC}_{(s_j,cyc_k,C_x)}$, and $\mathrm{LZC}_{(F_i,s_j,cyc_k,C_x)}$.

Table 1 lists all the main notations we use in this paper.

Table 1. A list of main notations defined in this paper.

Notation	Description
SS	The time series of sleep data
sc_i	The sleep cycle, $i \in \{1, n\}$
F_i	The four different frequency bands of EEG, i $\in \{Delat(D), Theta(T), Alpha(A), Beta(B)\}$
C_x	The channel of EEG, $x \in \{3, 4\}$
s_j	The j-th sleep stage, $j \in \{0, 1, 2, 3, 5\}$
cyc_k	The k-th sleep cycle, $k \in \{1, 2, 3, 4, 5\}$
E	The energy of frequency band
RP	The relative power
BSI	The brain symmetry index
SampEn	The sample entropy
DFA	The detrended fluctuation analysis
LZC	The Lempel–Ziv complexity
TST	Total Sleep Time, the interval between sleep onset and the end of sleep
SE	Sleep Efficiency, computed as the ratio between the hours of actual sleep and total time in bed
WASO	Wake After Sleep Onset, the proportion of awake stages during the whole sleep time
AT	Awake Times, the number of awakenings during the whole sleep time
SL	Sleep Latency, the time interval between going to bed and falling asleep
FF Trend 1	The sleep time of the last cycle is u minutes longer than that of the first cycle, we defined it as an upward trend, it is otherwise defined as a downward trend
TF Trend 1	The third cycle's deep sleep time (i.e., $s3$ and $s4$ stages) is v minutes longer than that of the first cycle, we defined it as an upward trend, it is otherwise defined as a downward trend
SF Trend 1	The second cycle's deep sleep time is w minutes longer than that of the first cycle, we defined it as an upward trend, it is otherwise defined as a downward trend

5. Stroke Prediction Model

In this section, the details of the proposed stroke prediction model are described, and the overview of the model is shown in Figure 4.

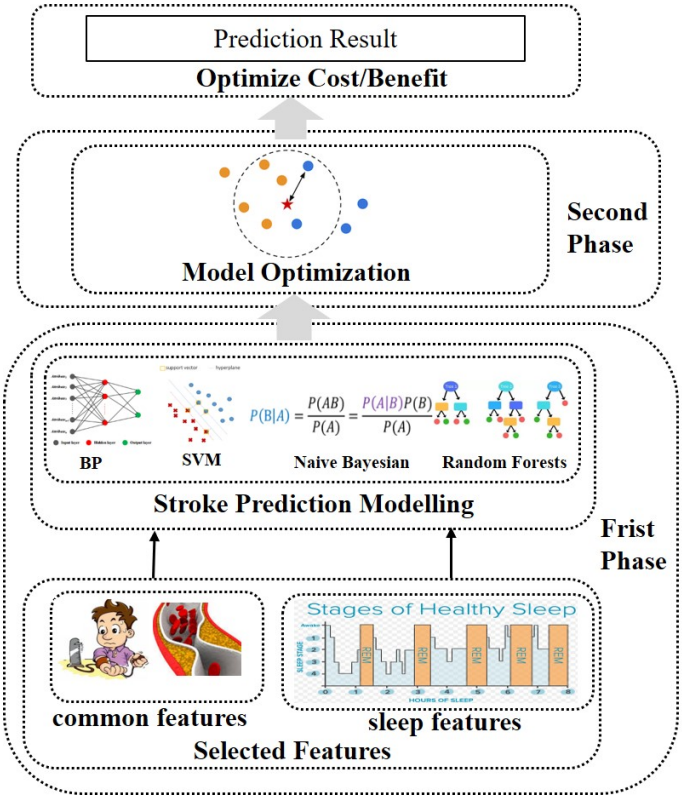

Figure 4. A general overview of the proposed two-step stroke prediction model.

5.1. Stroke Prediction Modelling

Based on the clinical features and the sleep-related features, four different classifiers are employed to construct the stroke prediction model, including a Support Vector Machine

(SVM), Random Forests (RF), a Back Propagation Neural Network (BPNN), Naive Bayesian (NB) and Logistic Regression (LR). Particularly, we utilize the 10-fold cross-validation technique to obtain more robust classification results [77].

SVM is one of the most popular classification algorithms and has been widely adopted to solve problems like regression and classification [78]. In this study, predicting stroke is a typical binary classification problem. SVM projects the input instances into a high-dimensional feature space, and then discriminates patients and healthy subjects by creating a hyperplane.

RF is a typical ensemble learning model for classification and regression tasks. It first constructs a set of decision trees at the training stage and then outputs a classification result that is either the mode of the classes or the mean prediction results of each individual tree [79].

BPNN is a common method for clustering or classifying, which is especially suitable for nonlinear systems. A typical BPNN consists of several layers of neurons and is characterized by network interconnection geometry, node characteristics, and the transfer functions [80].

NB is a Bayes' theorem-based probability classifier, which considers that each feature contributes independently to the probability for any possible correlations [1].

LR is a statistical analysis technique used to predict a binary outcome by evaluating the relationship between a set of independent predictor variables [81].

5.2. Model Optimization

For a classifier, the discrimination threshold is usually set to 0.5. If the outputted probability is larger than 0.5, the subject should be classified as a positive case, otherwise it should be labeled as a negative case. Undoubtedly, misclassification would lead to certain loss to the subjects. As for ischemia prediction, it will incur severe consequences if we classify a subject who is about to suffer from stroke as a healthy person. Thus, we should keep the false negative rate of the classification model as low as possible. At the same time, slight increase in the false positive rate of the classification model will not result in serious consequences. Therefore, we propose a pre-selection model to optimize the proposed stroke prediction model, in order to keep the false negative rate relatively low with the false positive rate not significantly increased.

In this paper, two different training datasets are utilized to optimize the model. Concretely, the first dataset is comprised of stroke patients (denoted as PartS) who suffered from stroke before and after the time of the SHHS-1 study. The second dataset consists of healthy subjects (denoted as PartH) who had no stroke throughout the whole data acquisition process. Next, based on the aforementioned two datasets, we construct a pre-selection model to determine which kind of threshold, i.e., a normal threshold or an adaptive threshold, should be employed when classifying each instance. In particular, we choose to use the k-NN classification algorithm to build the pre-selection model which classifies instances based on their k-closest neighbors' labels. Specifically, for each data instance, we first compute the distance between itself and all the instances in PartH and PartS, and then select the top-k nearest neighbors accordingly; if the ratio of instances from PartS is larger than a given threshold, we do have reason to believe that the instance is likely to suffer from stroke soon. Therefore, to make the classification model more sensitive to this kind of instance, an adaptive threshold smaller than 0.5 should be used. For other instances, we still use a normal threshold, i.e., 0.5. For example, if subject i's probability generated by the first phase model is smaller than 0.5, they would be classified to health person. However, in the pre-selection model, the ith instance is much closer to PartS, so we should reduce the threshold in order to classify it into potential stroke patients. Figure 5 shows the core idea of our KNN-based pre-selection model, where the k is set to 7, 11 and 15, respectively.

The details of the proposed model are described in Algorithm 1. Particularly, if a person is closer to PartS, we will reduce the threshold to a new threshold t that is smaller

than 0.5 accordingly. If the output of the model is greater than the new threshold t, we will classify the subjects into the Stroke class, instead of the Healthy class. Specifically, the cosine distance is used in the KNN model, which is in line with the general experience.

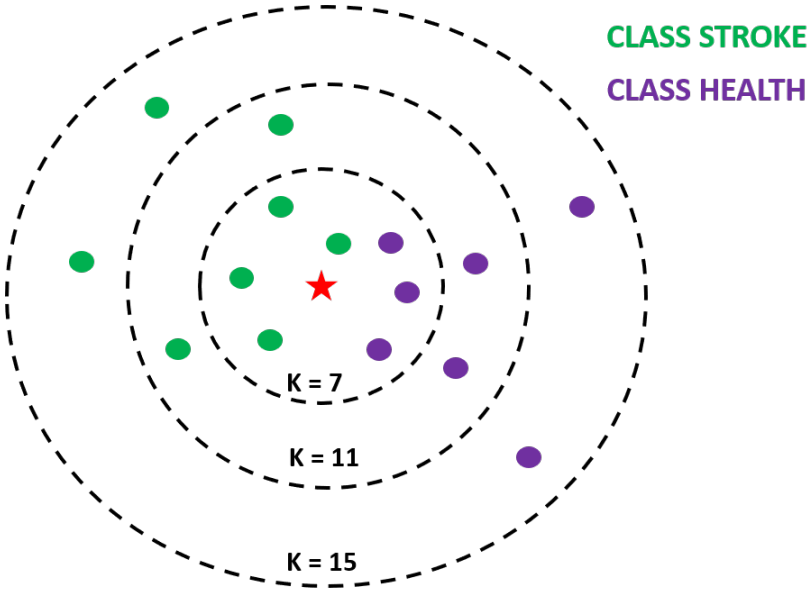

Figure 5. A pre-selection model based on kNN.

Algorithm 1 Pre-selection Model based on KNN.

Input:

p_i: subject i's probability distribution

n_test: a number of subjects in the testing dataset

t: the prediction threshold

subject_i: the i-th subject in the testing dataset

Output:

prediction result of the instance: 0 or 1

select a portion of data from each PartS and PartH;
for $i = 0; i < n\_test; i++$ **do**
  calculate the distance between i and samples in PartS and PartH;
  arrange results in ascending order;
  dataset_k $\leftarrow$ the top k closest samples;
  if sizeof(dataset_k $\cap$ PartS) $>$ k/2 and $p_i \geq t$ **then**
    label subject i as 1 (denotes ischemia);
  else
    if $p_i \geq 0.5$ **then**
      label subject i as 1 (denotes ischemia);
    else
      label subject i as 0 (denotes health);
    end if
  end if
end for

6. Experiment Evaluation

In this section, we will present the evaluation results of the proposed prediction model.

6.1. Experimental Setup

A dataset published by the Sleep Heart Health Study (SHHS) is used in this study. As a prospective multicenter cohort study, SHHS aims to study the subtle relationship between cardiovascular diseases and sleep-disordered breathing in the US. The participants were recruited from 9 epidemiological studies [82] enrolling more than 6600 adults aged 40 and over, who were asked to wear a home PSG device to record the occurrences of obstructive sleep apneas and detected risk factors for typical cardiovascular events (such as stroke and myocardial infarction) [82,83].

In the SHHS study, polysomnograms were collected in the participants' homes by trained technicians in unattended sittings [50]. The obtained data contain C4/A1and C3/A2 EEGs sampled at 125 Hz, right and left electrooculograms (EOGs) sampled at 50 Hz, a bipolar sub-mental electromyogram (EMG) sampled at 125 Hz, ambient light, body position, etc. In this study, two EEG channels (C3 and C4 channels) as well as the sleep stages (labeled every 30-s epoch) are used to construct the proposed ischemic stroke prediction model.

In this study, our analytical samples only included SHHS participants who completed PSG both in SHHS-1 and SHHS-2. The subjects are further categorized into four types according to the appearance of ischemic stroke. Specifically, the time when the SHHS-1 study was conducted is referred to as a separating timestamp. Type I subjects have stroke before the timestamp but have no stroke afterwards. Type II subjects have no stroke before the timestamp but have stroke afterwards. Type III subjects have stroke both before and after the timestamp. Type IV subjects have no stroke.

Specifically, to build the basic prediction model, we chose 179 subjects, which included 20 Type II instances (the time between their first ischemic stroke and the baseline ranging from 4 days to 377 days, with a mean value of 196 days) and 159 Type IV instances. In particular, the 20 positive samples were chosen based on the onset time of ischemic stroke, i.e., within one year, as the symptoms of a stroke are difficult to detect if the time exceeds one year. Moreover, considering that the classification performance would be less than ideal [84,85] if the ratio of positive to negative samples is too unbalanced, we chose 159 normal controls without medical history of cerebral vascular disease. Meanwhile, to build the pre-selection model, we chose another 57 subjects, which consisted of 27 Type III instances (formed the PartS in Algorithm 1) and 30 Type IV instances (formed the PartH in Algorithm 1).

We extracted the features described in [7] and used them as the feature set for the baseline method and compared it with the prediction model constructed using both these clinical features and the features that we extracted in Section 4. In the experiment, based on the empirical results, we set u = 42 min, v = 30 min, w = 4 min.

Statistical-based feature selection is the process of reducing the number of input variables when developing a predictive model. It works by calculating a test statistic to evaluate the relationship between each input variable and the target variable in order to improve the performance of the model. It calculates a probability value (p-value) to determine the significance of a feature based on the predefined significance level, which is typically set to 0.05. Specifically, in our experiment we used the Student t-test for calculating significant difference to select features.

The evaluation metrics we used include a Precision Recall Curve (PRC) [86], precision, and recall. A good prediction model should have both high Precision and Recall.

6.2. Evaluation Results

6.2.1. Sleep Structure Related Features

A summary of the sleep structure features extracted in this study is shown in Table 2. In addition to the mean value and standard deviation, the *p*-value of each feature is also calculated. Specifically, we can obtain the following observations according to Table 2.

Table 2. A summary of the sleep structure related features.

Feature	Type II	Type IV	*p*-Value
TST	668.7 ± 183.48	738.51 ± 125.59	0.0282
SE	68.31 ± 16.36	74.55 ± 11.8	0.0351
AT	32.65 ± 20.99	26.88 ± 11.52	0.0768
SL	88.8 ± 72.3	105.04 ± 80.05	0.389
REM-SL	354.55 ± 285.32	317.15 ± 173.14	0.419
WASO	23.35 ± 19	14.92 ± 13.87	0.0153
FF Trend 1	-17.1 ± 30.03	-37.93 ± 44.28	0.0425
TF Trend 1	-12.3 ± 26.21	-38.07 ± 44.71	0.0126
SF Trend 1	-3.15 ± 26.6	14.4 ± 45.72	0.2819

These features are shown as: mean $\pm$ std.

The analysis for the means and standard deviations of results (i.e., TST and SE: the total sleep time and sleep efficiency of Type II are less than those for Type IV) indicates that the sleep quality of potential patients is obviously worse than that of the ordinary people, meanwhile the fact that WASO of Type II is bigger than Type IV also confirms this point. Moreover, two of the remaining features (i.e., FF trend 1 and TF trend 1) have relatively small *p*-values ($p < 0.05$), indicating that between these potential ischemia and healthy people there is significant difference. According to Table 1, there are significant differences between some of the sleep structure related features, which indicates that the sleep quality of Type II subjects is worse than that of healthy people, e.g., the former ones have shorter sleep time and lower sleep efficiency. However, another four features' *p*-value (such as SL and REM-SL) are larger than 0.05, indicating that these features cannot help us to solve the problem.

6.2.2. EEG Sleep Relate Features

A summary of the sleep EEG related features extracted in this study, which include relative power, DAR, TBR, DTABR, TDABR, BSI, SampEn, LZC and DFA, is shown in this section. Significantly, we used two different EEG channels (C3 and C4), which were the interhemispheric electrode pairs. Therefore, each feature has a pair of values.

All the relative power features of four different frequency bands (delta, theta, alpha, beta) are summarized in Table 3.

From Table 3, we can see that there is no significant difference, which means that we cannot differentiate Type II and Type IV. However, as aforementioned, people usually pass through five sleep stages: 0, 1, 2, 3 (includes 4) and 5. In order to study the power ratio changes during different sleep stages, we calculated the relative power features of different frequency bands during different sleep stages, as shown in Tables 4–7.

Table 4 shows the comparison of the relative power of frequency band F_D on EEG channel C_x during the j-th sleep stage for ischemic stroke patients and control subjects, where $x \in \{3, 4\}$, and $j \in \{0, 1, 2, 3, 5\}$.

Table 3. A summary of different frequency bands' power ratio features.

Feature	Type II	Type IV	*p*-Value
$RP_{(F_D,C_3)}$	0.761 ± 0.188	0.795 ± 0.087	0.156
$RP_{(F_T,C_3)}$	0.089 ± 0.040	0.094 ± 0.039	0.569
$RP_{(F_A,C_3)}$	0.073 ± 0.173	0.044 ± 0.033	0.051
$RP_{(F_B,C_3)}$	0.076 ± 0.048	0.068 ± 0.043	0.341
$RP_{(F_D,C_4)}$	0.793 ± 0.083	0.784 ± 0.109	0.748
$RP_{(F_T,C_4)}$	0.096 ± 0.044	0.093 ± 0.035	0.653
$RP_{(F_A,C_4)}$	0.045 ± 0.057	0.05 ± 0.057	0.762
$RP_{(F_B,C_4)}$	0.065 ± 0.025	0.073 ± 0.055	0.528

These features are shown as: mean $\pm$ std.

Table 4. A summary of delta power ratio features during different sleep stages.

Feature	Type II	Type IV	*p*-Value
$RP_{(F_D,s_0,C_3)}$	0.233 ± 0.145	0.178 ± 0.109	0.043
$RP_{(F_D,s_1,C_3)}$	0.020 ± 0.027	0.014 ± 0.014	0.085
$RP_{(F_D,s_2,C_3)}$	0.296 ± 0.097	0.304 ± 0.111	0.754
$RP_{(F_D,s_3,C_3)}$	0.120 ± 0.127	0.188 ± 0.131	0.030
$RP_{(F_D,s_5,C_3)}$	0.042 ± 0.033	0.058 ± 0.034	0.053
$RP_{(F_D,s_0,C_4)}$	0.203 ± 0.144	0.164 ± 0.104	0.137
$RP_{(F_D,s_1,C_4)}$	0.019 ± 0.023	0.014 ± 0.014	0.155
$RP_{(F_D,s_2,C_4)}$	0.359 ± 0.130	0.314 ± 0.111	0.098
$RP_{(F_D,s_3,C_4)}$	0.130 ± 0.146	0.200 ± 0.134	0.032
$RP_{(F_D,s_5,C_4)}$	0.048 ± 0.037	0.055 ± 0.032	0.344

These features are shown as: mean $\pm$ std.

Table 5 shows the comparison of the relative power of the frequency band F_T on the EEG channel C_x during the j-th sleep stage for ischemic stroke patients and control subjects, where $x \in \{3,4\}$, and $j \in \{0,1,2,3,5\}$.

Table 5. A summary of theta power ratio features during different sleep stages.

Feature	Type II	Type IV	*p*-Value
$RP_{(F_T,s_0,C_3)}$	0.017 ± 0.011	0.009 ± 0.007	2.48×10^{-4}
$RP_{(F_T,s_1,C_3)}$	0.002 ± 0.002	0.002 ± 0.002	0.988
$RP_{(F_T,s_2,C_3)}$	0.047 ± 0.025	0.048 ± 0.024	0.954
$RP_{(F_T,s_3,C_3)}$	0.009 ± 0.01	0.018 ± 0.015	0.020
$RP_{(F_T,s_5,C_3)}$	0.006 ± 0.004	0.008 ± 0.005	0.052
$RP_{(F_T,s_0,C_4)}$	0.014 ± 0.008	0.009 ± 0.003	0.007
$RP_{(F_T,s_1,C_4)}$	0.002 ± 0.002	0.002 ± 0.001	0.419
$RP_{(F_T,s_2,C_4)}$	0.055 ± 0.031	0.048 ± 0.024	0.263
$RP_{(F_T,s_3,C_4)}$	0.011 ± 0.012	0.018 ± 0.014	0.047
$RP_{(F_T,s_5,C_4)}$	0.007 ± 0.005	0.008 ± 0.004	0.396

These features are shown as: mean $\pm$ std.

Table 6 shows the comparison of the relative power of frequency band F_A on EEG channel C_x during the j-th sleep stage for ischemic stroke patients and control subjects, where $x \in \{3,4\}$, and $j \in \{0,1,2,3,5\}$.

Table 7 shows the comparison of the relative power of frequency band F_B on the EEG channel C_x during the j-th sleep stage for ischemic stroke patients and control subjects, where $x \in \{3,4\}$, and $j \in \{0,1,2,3,5\}$.

Different from the short-term EEG test in clinical settings, this study used overnight EEG data which may contain more abundant information. As we mentioned in Figure 1, changes in EEG morphology and frequency correlate with reductions in CBF [28]. Particularly, as indicated by the ischemic threshold in Figure 1, the slower frequencies (4~7 Hz) of ischemic patients should gradually increase, which is in accordance with the results in Table 4. For example, the mean value and standard deviation of features (including $RP_{(F_T,s_0,C_3)}$, $RP_{(F_T,s_0,C_4)}$, $RP_{(F_T,s_0,cyc_3,C_3)}$, $RP_{(F_T,s_0,cyc_3,C_4)}$) measuring the sleep quality

of stroke patients (i.e., Type II) are obviously larger than that of healthy individuals (i.e., Type IV).

Table 6. A summary of alpha power ratio features during different sleep stages.

Feature	Type II	Type IV	*p*-Value
$RP_{(F_A,s_0,C_3)}$	0.009 ± 0.010	0.007 ± 0.010	0.326
$RP_{(F_A,s_1,C_3)}$	0.0007 ± 0.0006	0.0009 ± 0.001	0.350
$RP_{(F_A,s_2,C_3)}$	0.039 ± 0.129	0.016 ± 0.010	0.022
$RP_{(F_A,s_3,C_3)}$	0.001 ± 0.001	0.003 ± 0.003	0.011
$RP_{(F_A,s_5,C_3)}$	0.001 ± 0.001	0.002 ± 0.002	0.104
$RP_{(F_A,s_0,C_4)}$	0.008 ± 0.004	0.007 ± 0.010	0.740
$RP_{(F_A,s_1,C_4)}$	0.001 ± 0.002	0.001 ± 0.001	0.532
$RP_{(F_A,s_2,C_4)}$	0.017 ± 0.024	0.019 ± 0.021	0.669
$RP_{(F_A,s_3,C_4)}$	0.001 ± 0.001	0.004 ± 0.007	0.109
$RP_{(F_A,s_5,C_4)}$	0.005 ± 0.014	0.003 ± 0.010	0.648

These features are shown as: mean $\pm$ std.

Table 7. A summary of beta power ratio features during different sleep stages.

Feature	Type II	Type IV	*p*-Value
$RP_{(F_B,s_0,C_3)}$	0.015 ± 0.016	0.009 ± 0.011	0.070
$RP_{(F_B,s_1,C_3)}$	0.0004 ± 0.0004	0.0004 ± 0.0006	0.786
$RP_{(F_B,s_2,C_3)}$	0.007 ± 0.016	0.004 ± 0.004	0.017
$RP_{(F_B,s_3,C_3)}$	0.0005 ± 0.0009	0.0006 ± 0.001	0.612
$RP_{(F_B,s_5,C_3)}$	0.0007 ± 0.0005	0.001 ± 0.001	0.042
$RP_{(F_B,s_0,C_4)}$	0.009 ± 0.007	0.009 ± 0.009	0.681
$RP_{(F_B,s_1,C_4)}$	0.0004 ± 0.0005	0.0005 ± 0.0005	0.603
$RP_{(F_B,s_2,C_4)}$	0.004 ± 0.003	0.005 ± 0.009	0.643
$RP_{(F_B,s_3,C_4)}$	0.0005 ± 0.0008	0.001 ± 0.005	0.594
$RP_{(F_B,s_5,C_4)}$	0.0008 ± 0.0006	0.001 ± 0.003	0.178

These features are shown as: mean $\pm$ std.

Moreover, to further characterize the differences between Type II and Type IV, we also calculated the relative power of different frequency bands during different sleep stages and different sleep cycles (from sleep cycle 1 to sleep cycle 5). However, as each frequency band has more than 50 features, we only list features of frequency band theta, due to the limited space available.

Table 8 shows the comparison of the relative power of frequency band F_T on the EEG channel C_x during the j-th sleep stage of the k-th sleep cycle for ischemic stroke patients and control subjects, where $x \in \{3,4\}$, $j \in \{0,1,2,3,5\}$, and $k \in \{1,2,3,4,5\}$. Meanwhile we also calculated the corresponding features of the other three frequency bands (i.e., delta, alpha, beta); they are far too numerous to list each of them individually.

Table 8. A summary of theta power ratio features during different sleep stages.

Feature	Type II	Type IV	*p*-Value
$RP_{(F_T,s_3,cyc_1,C_3)}$	0.016 ± 0.016	0.029 ± 0.022	0.015
$RP_{(F_T,s_5,cyc_1,C_3)}$	0.012 ± 0.012	0.008 ± 0.008	0.038
$RP_{(F_T,s_0,cyc_2,C_3)}$	0.016 ± 0.016	0.008 ± 0.009	0.003
$RP_{(F_T,s_0,cyc_3,C_3)}$	0.020 ± 0.014	0.008 ± 0.008	3.062×10^{-5}
$RP_{(F_T,s_1,cyc_3,C_3)}$	0.007 ± 0.011	0.002 ± 0.002	1.68×10^{-4}
$RP_{(F_T,s_3,cyc_1,C_4)}$	0.018 ± 0.018	0.029 ± 0.022	0.033
$RP_{(F_T,s_5,cyc_1,C_4)}$	0.013 ± 0.014	0.007 ± 0.006	0.003
$RP_{(F_T,s_0,cyc_2,C_4)}$	0.014 ± 0.013	0.008 ± 0.009	0.016
$RP_{(F_T,s_0,cyc_3,C_4)}$	0.021 ± 0.015	0.009 ± 0.009	1.71×10^{-5}
$RP_{(F_T,s_1,cyc_3,C_4)}$	0.006 ± 0.01	0.002 ± 0.002	2.19×10^{-4}

These features are shown as: mean $\pm$ std.

As shown in Table 8, we found that during sleep stage 0 of the 3rd sleep cycle there is a significant difference between the two categories. Although there are similar phenomena

in other waveforms (including delta, alpha, beta), it is more obvious in the theta band. According to data analysis, we found that the third sleep cycles are usually between 1: 00 a.m. and 3: 00 a.m. Therefore, we preliminary concluded that the theta waveform in sleep stage 0 of the third sleep cycle was a potential risk factor that might have been previously unexplored .

As aforementioned, we also extracted several other frequency domain features, including BSI, DAR, TBR, DTABR, TDABR, $\mathrm{BSI}_{(s_j,cyc_k,C_x)}$, $\mathrm{DAR}_{(F_i,s_j,cyc_k,C_x)}$, $\mathrm{TBR}_{(F_i,s_j,cyc_k,C_x)}$, $\mathrm{DTABR}_{(F_i,s_j,cyc_k,C_x)}$ and $\mathrm{TDABR}_{(F_i,s_j,cyc_k,C_x)}$. However, none of them had significant difference between subjects of Type II and Type IV. For example, the p-value of BSI is 0.902, indicating that subjects of Type II and Type IV have almost equal CBF between the left and right hemispheres.

Table 9 shows the comparison of the Sample Entropy of the EEG on channel C_x during the j-th sleep stage of the k-th sleep cycle for ischemic stroke patients and control subjects, where $x \in \{3,4\}$, j$\in \{0,1,2,3,5\}$, and $k \in \{1,2,3,4,5\}$.

Table 9. A summary of the Sample Entropy features during different sleep stages of different sleep cycles.

Feature	Type II	Type IV	p-Value
$\mathrm{SampEn}_{(s_1,cyc_1,C_3)}$	0.855 ± 0.619	1.079 ± 0.381	0.023
$\mathrm{SampEn}_{(s_3,cyc_1,C_3)}$	0.479 ± 0.348	0.617 ± 0.211	0.012
$\mathrm{SampEn}_{(s_0,cyc_3,C_3)}$	1.198 ± 0.167	0.884 ± 0.327	0.0001
$\mathrm{SampEn}_{(s_1,cyc_1,C_4)}$	0.737 ± 0.588	1.018 ± 0.353	0.002
$\mathrm{SampEn}_{(s_3,cyc_1,C_4)}$	0.418 ± 0.293	0.585 ± 0.201	0.001
$\mathrm{SampEn}_{(s_5,cyc_1,C_4)}$	0.832 ± 0.410	0.995 ± 0.241	0.010
$\mathrm{SampEn}_{(s_3,cyc_2,C_4)}$	0.366 ± 0.322	0.527 ± 0.262	0.019
$\mathrm{SampEn}_{(s_5,cyc_2,C_4)}$	0.794 ± 0.354	0.980 ± 0.275	0.011
$\mathrm{SampEn}_{(s_0,cyc_3,C_4)}$	1.052 ± 0.292	0.792 ± 0.270	0.0008

These features are shown as: mean $\pm$ std.

Similarly, we also calculated the sample entropy of frequency bands delta, theta, alpha and beta on channel C_x during the j-th sleep stage of the k-th sleep cycle for ischemic stroke patients and control subjects, as well as the LZC and DFA. For the same reason described above, we did not list them one by one. According to the results of nonlinear features, we found that the most significant differences between types II and IV are concentrated on sleep cycles 1 and 2. The reason might be that subjects of Type II might have difficulty in falling into a stable state during the first sleep cycle.

In order to acquire a better performance of classification and reduce the computing cost and the time delay in successive steps, we preserved the feature that made a significant contribution to the prediction. Specifically, we leverage information gain to measure the effectiveness of candidate features, i.e., a feature will be discarded if its information gain is smaller than 0.05. The input of the classifier is a vector consisting of the remaining features.

To sum up, we mainly used $RP_{(F_T,s_0,cyc_3,C_4)}$, $RP_{(F_D,s_0,cyc_3,C_3)}$, $RP_{(F_A,s_0,cyc_3,C_4)}$, $\mathrm{SampEn}_{(s_0,cyc_3,C_3)}$, $\mathrm{SampEn}_{(s_0,cyc_3,C_4)}$, $\mathrm{SampEn}_{(F_D,s_3,cyc_1,C_3)}$, $\mathrm{SampEn}_{(F_D,s_3,cyc_1,C_4)}$, $\mathrm{LZC}_{(F_D,s_2,cyc_1,C_3)}$, $\mathrm{LZC}_{(F_D,s_3,cyc_1,C_3)}$, $\mathrm{LZC}_{(F_D,s_3,cyc_1,C_4)}$, $\mathrm{LZC}_{(F_T,s_3,cyc_1,C_4)}$, $\mathrm{LZC}_{(F_A,s_1,cyc_1,C_4)}$, $\mathrm{DFA}_{(s_2,cyc_1,C_4)}$, WASO and TF trend 1 to distinguish Type II and Type IV subjects.

Tables 10 and 11 shows the clinical features and main polysomnogram features we used in experiments, respectively.

Table 10. List of the clinical features.

Feature	Description
SYST	Systolic blood pressure
DIAS	Diastolic blood pressure
HTNDerv	Hypertension
ParRpDiab	History of diabetes
AFIB	Atrial fibrillation or flutter
Lvh	Left ventricular hypertrophy
Chol	Cholesterol

Table 11. List of the polysomnogram features.

Feature	Description
$\text{SampEn}_{(s_0,cyc_3,C_3)}$	The sample entropy on EEG channel C_3 during the 0-th sleep stage of the 3-th sleep cycle
$\text{SampEn}_{(s_0,cyc_3,C_4)}$	The sample entropy on EEG channel C_4 during the 0-th sleep stage of the 3-th sleep cycle
$\text{SampEn}_{(F_D,s_3,cyc_1,C_3)}$	The sample entropy of frequency band F_D on EEG channel C_3 during the 3-th sleep stage of the 1-th sleep cycle
$\text{SampEn}_{(F_D,s_3,cyc_1,C_4)}$	The sample entropy of frequency band F_D on EEG channel C_4 during the 3-th sleep stage of the 1-th sleep cycle
$\text{LZC}_{(F_D,s_2,cyc_1,C_3)}$	The Lempel–Ziv complexity of frequency band F_D on EEG channel C_3 during the 2-th sleep stage of the 1-th sleep cycle
$\text{LZC}_{(F_D,s_3,cyc_1,C_3)}$	The Lempel–Ziv complexity of frequency band F_D on EEG channel C_3 during the 3-th sleep stage of the 1-th sleep cycle
$\text{LZC}_{(F_D,s_3,cyc_1,C_4)}$	The Lempel–Ziv complexity of frequency band F_D on EEG channel C_4 during the 3-th sleep stage of the 1-th sleep cycle
$\text{DFA}_{(s_2,cyc_1,C_4)}$	The detrended fluctuation analysis on EEG channel C_4 during the 2-th sleep stage of the 1-th sleep cycle
TF trend 1	The third cycle's deep sleep time (i.e., $s3$ and $s4$ stages) is v minutes longer than that of the first cycle, we defined it as an upward trend, it is otherwise defined as a downward trend.

Figure 6 shows a visualization in terms of several extracted features. With these views, we can find that the features of two types follow quite a different distribution. However, due to the overlap of feature distributions, we cannot distinguish a potential ischemic patient and a healthy person by using any single-feature. Therefore, we arrange all the useful attributes as a vector and then feed them into the prediction model.

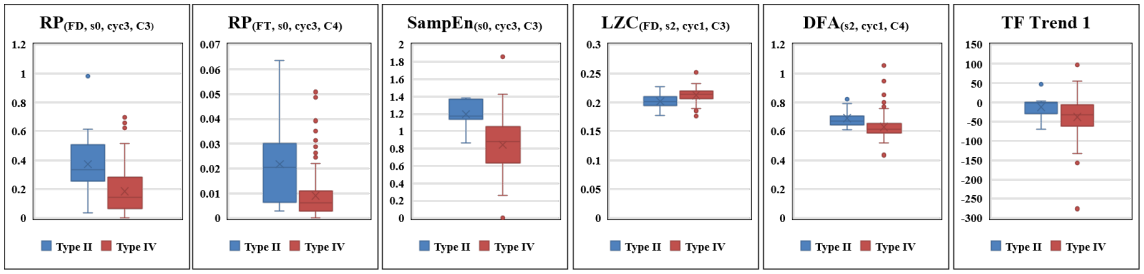

Figure 6. The box-plots of all features. The whiskers represent the smallest and largest observations, the edges of the box correspond to the lower and upper quartiles, the horizontal line indicates the median and the plus sign marks probable outliers.

6.2.3. Stroke Prediction Results

We use BPNN, SVM, NB, RF and LR as the classification algorithm, and adopted 10-fold cross validation to verify the correctness of the experimental results. The original positive and negative samples were randomly partitioned into 10 equal-sized subsamples, respectively. Then we fitted our model to a data set consisting of 9 of the original 10 parts (including 18 positive and 144 negative samples), and the remaining one subsample was used as the testing dataset. The cross-validation process is repeated 10 times, with each of the 10 subsamples used exactly once as the testing dataset. The 10 results were assembled

to produce a single estimation. As shown in Figure 7, the Precision-Recall (PR) curves indicate that the proposed model can keep a relatively high TPR and obtain a low FPR at the same time.

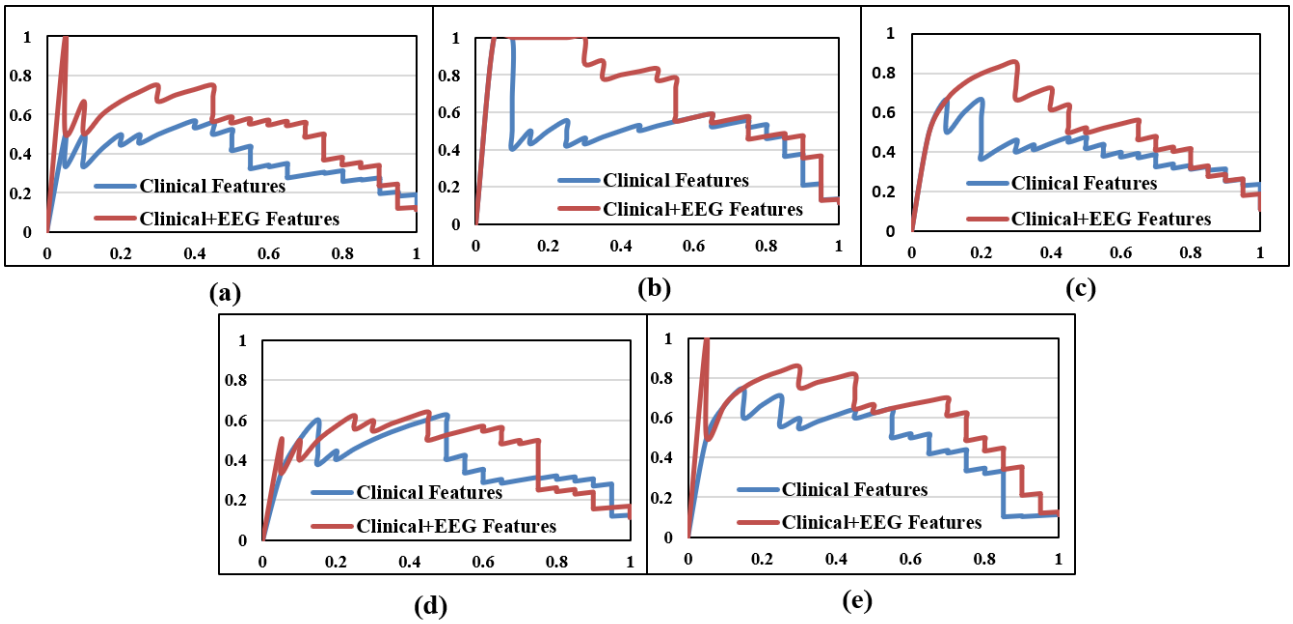

Figure 7. Precision Recall (PR) curves of the proposed features and the baseline features. (**a**) Back Propagation (BP) neural network models PR curve, (**b**) Support Vector Machine (SVM) models PR curve, (**c**) Random Forest models PR curve, (**d**) Naïve Bayesian models PR curve, (**e**) Logistic Regressio (LR) models PR curve.

Figure 7 shows that the combination of the baseline features and the proposed features is superior to the baseline features. Obviously, a good classification method should have both a high detection rate and low false alarm rate; therefore, the pair (Precision, Recall) is widely used to evaluate the performance of classification method. They are defined as:

$$precision = \frac{TP}{TP + FP}. \tag{7}$$

$$recall = \frac{TP}{TP + FN}. \tag{8}$$

The models are further evaluated regarding four metrics, namely, the mean precision, recall, and F-Measure. As shown in Figure 8, the models perform better when using feature combinations than using baseline features. Taking the SVM-based model as an example, the precision, recall and PRC of the SVM-based model were improved by 27%, 55% and 46%, respectively.

Specifically, in Figure 8, comparison between the TPR and the FPR for both the selected algorithms allows the observation that SVM (parameters: k = Polynomial and c = 0.8) and LR achieved better performance (higher TPR and lower FPR) than the other three models. Therefore, we built our basic models using SVM and LR, and then further optimized them using the pre-selection model (parameter: K = 7 in the pre-selection model). In particular, through testing different adaptive thresholds in algorithm 1, we found that the results of SVM were superior to LR, with an acceptance threshold equal to 0.2, depending on the empirical results. In terms of true positive rate and false positive rate, the performance of the optimized prediction model is shown in Figure 9.

Accordingly, compared to the basic SVM, the SVM+pre-selection model achieves better performance, where the true positive rate was increased by 21%. Specifically, as shown in Figure 10, the optimized prediction model can predict 17 of the 20 Type II participants successfully, and only 10 of the 159 Type IV participants are labeled wrongly, while the

two values of the basic model are 14 and 8. Accordingly, we can find that the optimized model will help an additional three Type II subjects to receive timely medical treatment, although it will also trouble the other two type IV subjects unnecessarily. Nevertheless, we can conclude that the optimized model has better performance than the basic model.

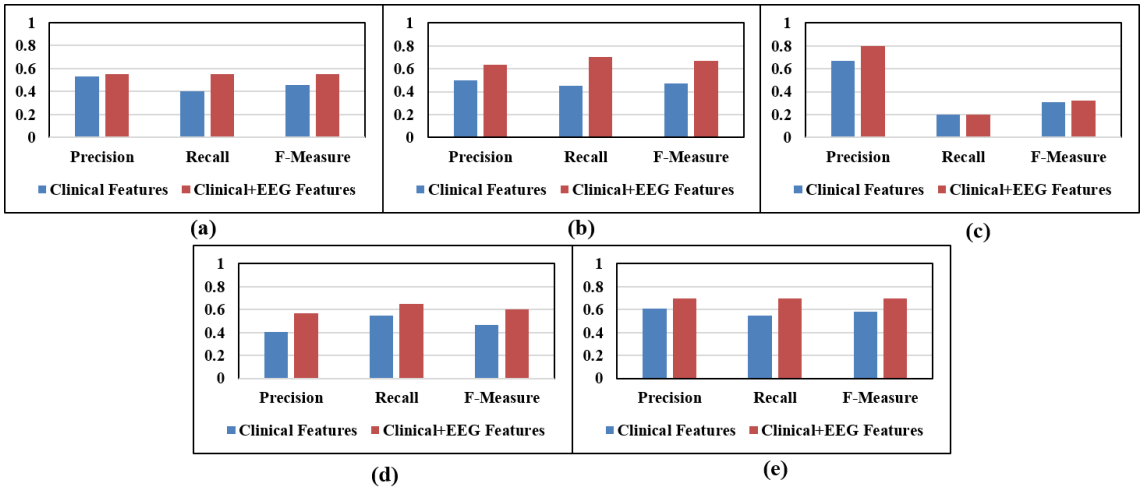

Figure 8. Performance of the proposed model. (**a**) BP neural network, (**b**) SVM, (**c**) Random Forests, and (**d**) Naive Bayesian, (**e**) Logistic Regression.

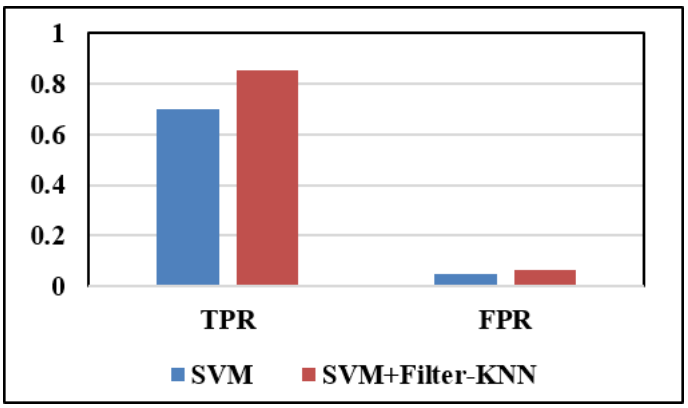

Figure 9. Performance of the proposed method.

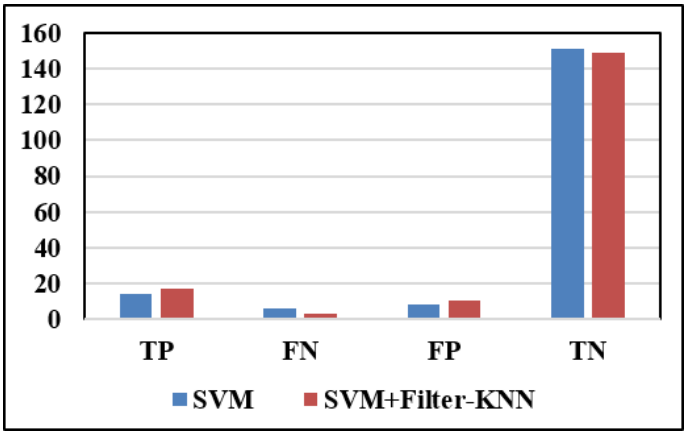

Figure 10. Comparison of the two models in TP, FN, FP, TN.

6.3. Limitations

The main strength of our study is that we put forward a novel prediction method by exploring sleep related features. However, some limitations of this study should be pointed out.

Firstly, we evaluate the framework using a real polysomnogram dataset that contains 20 stroke patients and 159 healthy individuals. Compared to the percentage of ischemic stroke in the real world, the number of health samples examined in experiments is very small. For example, the prevalence of stroke in China in 2013 was 1114.8 per 100,000 people [87]. We could not obtain enough information of normal controls to further estimate our prediction model and polysomnogram features. Therefore, our experimental setup cannot be indicative of the real situation of ischemia and might have influenced the performance of our model. Nevertheless, by adopting reasonable indicators (i.e., precision, recall and area under the precision–recall curve), we validated the predictive effects of our model and features in imbalanced data and proved the rationality and effectiveness to some extent.

Secondly, the participants in SHHS were asked to wear a home PSG device to record the occurrences of obstructive sleep apneas (OSA). OSA elicit increases in sympathetic nerve activity with higher ischemic stroke incidents, meanwhile patients with OSA manifest marked increases in blood pressure during sleep [88]. Thus, EEG features might have shown hyperactivity and sympathetic function resulting hypertension or other clinical factors and might not have been totally caused by cerebral ischemia. The correlation between clinical features and EEG features was analyzed by the correlation coefficient and it was found that EEG features were moderately or weakly associated with clinical features. However, the limited number of samples may not be able to fully explore the correlation among them. Further comprehensive association analyses and functional experiments are required in the future.

Thirdly, CBF is the major cause of ischemic stroke. Although EEG abnormalities are closely tied to CBF, it directly affects the reliability and accuracy of the result because data on CBF levels are not available in our study. Based on a suitable public and real data set, developing a more robust prediction system to study the relationship between stroke and CBF will be one of our future works.

7. Conclusions and Future Work

A predictive model relies on the association relationship among multi-dimension features, and thus feature extraction plays an important role in improving the ischemic stroke identification performance. In this paper, we not only extract clinical features from clinical history, demographic information, physical and biomedical measurements but also extract features from a polysomnogram to reflect the change pattern of sleep. Experimental results show that the extracted features can characterize sleep patterns more comprehensively, and hence generates more information for classification. However, it does not imply that all the extracted features are useful, as some of them might be redundant and even irrelevant to the classification problem. Therefore, we used a statistical test and information gain to select the extracted features, which include clinical features, sleep structure related features and EEG sleep related features. Finally, we developed an ischemic stroke prediction model that consists of two steps: the first one uses a machine learning model to make the basic prediction, and the second one further optimizes the prediction result by controlling the false negative rate.

Experimental results proved that the combined characteristics are better than using clinical features alone. At the sleep structured level, the analysis on the result indicates that would-be patients have shorter sleep time and lower sleep efficiency, as well as increasing in wake after sleep onset. In sleep trends, a prolonged lighter sleep stages were found in most experiment groups, and the control group gained more stable sleep-wake stages. It indicates that the patients have significantly reduced sleep efficiency. From the point of view of the electroencephalogram, changes in EEG morphology and frequency correlate with reductions in CBF and stroke symptoms can result from inadequate CBF.

In addition, more and more data from evidence-based medicine demonstrate a higher risk of stroke in the early morning hours (03:00 a.m. to 06:00 a.m.) [89]. As shown in Table 7, we found that during light sleep of the third sleep cycle (according to data analysis the third sleep cycle is usually between 2:00 a.m. and 4:00 a.m.), there is a significant difference between the controls and the patients. Although there are similar phenomena in other waveforms (including delta, alpha, beta), it is more obvious in the theta band. Experimental results also show that potential ischemic patients tend to have more slow waves in most frequency domain features and nonlinear features, which is just the opposite of the control subjects.

As future work, we plan to extend this work in two directions. First, feature extraction plays an important role in machine learning; thus, by considering other features (i.e., fractal dimension [90] and Lyapunov exponent [91]) we will try to obtain better prediction performance. Second, deep learning is achieving amazing results in various research areas; therefore, we will explore deep learning models for the construction of more effective stroke prediction models.

Author Contributions: Conceptualization, Z.W.; data curation, J.X.; formal analysis, J.X. and Z.W.; funding acquisition, Z.W., Z.Y. and B.G.; methodology, J.X. and Z.W.; project administration, J.X.; supervision, Z.W., Z.Y., B.G. and X.Z.; writing—original draft, J.X.; writing—review and editing, Z.W., Z.Y. and B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2016YFB1001400), and the National Natural Science Foundation of China (No. 61332005, 61772428), the Innovative Talents Promotion Program of Shaanxi Province (No. 2018KJXX-011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Available online: http://en.wikipedia.org (accessed on 10 January 2021).
2. Available online: http://www.strokecenter.org (accessed on 10 January 2021).
3. Bao, M.H.; Szeto, V.; Yang, B.B.; Zhu, S.Z.; Sun, H.S.; Feng, Z.P. Long non-coding RNAs in ischemic stroke. *Cell Death Dis.* **2018**, *9*, 281. [CrossRef] [PubMed]
4. Available online: http://www.stroke.org (accessed on 10 January 2021).
5. Available online: https://www.nhlbi.nih.gov/health-topics/stroke (accessed on 10 January 2021).
6. Sedghi, E.; Weber, J.H.; Thomo, A.; Bibok, M.; Penn, A.M. Mining clinical text for stroke prediction. *Netw. Model. Anal. Health Informat. Bioinform.* **2015**, *4*, 16. [CrossRef]
7. Lumley, T.; Kronmal, R.A.; Cushman, M.; Manolio, T.A.; Goldstein, S. A stroke prediction score in the elderly: Validation and Web-based application. *J. Clin. Epidemiol.* **2002**, *55*, 129–136. [CrossRef]
8. Longstreth, W.T., Jr.; Bernick, C.; Fitzpatrick, A.; Cushman, M.; Knepper, L.; Lima, J.; Furberg, C.D. Frequency and predictors of stroke death in 5888 participants in the Cardiovascular Health Study. *Neurology* **2001**, *56*, 368–375. [CrossRef] [PubMed]
9. Manolio, T.A.; Kronmal, R.A.; Burke, G.L.; O'Leary, D.H.; Price, T.R. Short-term predictors of incident stroke in older adults. The Cardiovascular Health Study. *Stroke* **1996**, *27*, 1479. [CrossRef] [PubMed]
10. McGinn, A.P.; Kaplan, R.C.; Verghese, J.; Rosenbaum, D.M.; Psaty, B.M.; Baird, A.E.; Lynch, J.K.; Wolf, P.A.; Kooperberg, C.; Larson, J.C.; et al. Walking speed and risk of incident ischemic stroke among postmenopausal women. *Stroke* **2008**, *39*, 1233–1239. [CrossRef]
11. Sajjadi, M.; Karami, M.; Amirfattahi, R.; Bateni, V.; Ahamadzadeh, M.R.; Ebrahimi, B. A promising method of enhancement for early detection of ischemic stroke. *J. Res. Med. Sci. Off. J. Isfahan Univ. Med. Sci.* **2012**, *17*, 843–849.
12. Lau, J.K.; Lowres, N.; Neubeck, L. iPhone ECG application for community screening to detect silent atrial fibrillation: A novel technology to prevent stroke. *Int. J. Cardiol.* **2013**, *165*, 193–194. [CrossRef] [PubMed]
13. Lowres, N.; Neubeck, L.; Salkeld, G.; Krass, I.; McLachlan, A.J.; Redfern, J.; Freedman, S.B. Feasibility and cost-effectiveness of stroke prevention through community screening for atrial fibrillation using iPhone ECG in pharmacies. The SEARCH-AF study. *Thromb. Haemost.* **2014**, *111*, 1167–1176. [CrossRef]
14. Shanthi, A.S.; Karthikeyan, M. Support Vector Machine for MRI Stroke Classfication. *Int. J. Comput. Sci. Eng.* **2014**, *6*, 156.

15. Khosla, A.; Cao, Y.; Lin, C.C.Y.; Chiu, H.K.; Hu, J.; Lee, H. An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–28 July 2010; pp. 183–192.
16. Fried, L.P.; Borhani, N.O.; Enright, P.; Furberg, C.D.; Gardin, J.M.; Kronmal, R.A.; MPH for the Cardiovascular Health Study Research Group. The Cardiovascular Health Study: Design and rationale. *Ann. Epidemiol.* **1991**, *1*, 263. [CrossRef]
17. Othman, M.; Kasabov, N.; Tu, E.; Feigin, V.; Krishnamurthi, R.; Hou, Z.; Hu, J. Improved predictive personalized modeling with the use of Spiking Neural Network system and a case study on stroke occurrences data. In Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 6–11 July 2014; pp. 3197–3204.
18. Elliott, W.J. Circadian variation in the timing of stroke onset: A metaanalysis. *Stroke* **1998**, *29*, 992–996. [CrossRef] [PubMed]
19. Marler, J.R.; Price, T.R.; Clark, G.L.; Muller, J.E.; Robertson, T.; Mohr, J.P.; Foulkes, M.A. Morning increase in onset of ischemic stroke. *Stroke* **1989**, *20*, 473–476. [CrossRef]
20. Marsh, E.E.; Biller, J.; Adams, H.P.; Marler, J.R.; Hulbert, J.R.; Love, B.B.; Gordon, D.L. Circadian variation in onset of acute ischemic stroke. *Arch. Neurol.* **1990**, *47*, 1178–1180. [CrossRef]
21. Körner, E.; Flooh, E.; Reinhart, B.; Wolf, R.; Ott, E.; Krenn, W.; Lechner, H. Sleep alterations in ischemic stroke. *Eur. Neurol.* **1986**, *25* (Suppl. 2), 104–110. [CrossRef]
22. Koo, D.L.; Nam, H.; Thomas, R.J.; Yun, C.H. Sleep disturbances as a risk factor for stroke. *J. Stroke* **2018**, *20*, 12. [CrossRef]
23. Murri, L.; Gori, S.; Massetani, R.; Bonanni, E.; Marcella, F.; Milani, S. Evaluation of acute ischemic stroke using quantitative EEG: A comparison with conventional EEG and CT scan. *Neurophysiol. Clin. Neurophysiol.* **1998**, *28*, 249–257. [CrossRef]
24. Ma, C.; Pavlova, M.; Liu, Y.; Liu, Y.; Huangfu, C.; Wu, S.; Gao, X. Probable REM sleep behavior disorder and risk of stroke: A prospective study. *Neurology* **2017**, *88*, 1849. [CrossRef] [PubMed]
25. Finnigan, S.; Wong, A.; Read, S. Defining abnormal slow EEG activity in acute ischaemic stroke: Delta/alpha ratio as an optimal QEEG index. *Clin. Neurophysiol.* **2016**, *127*, 1452–1459. [CrossRef] [PubMed]
26. Foreman, B.; Claassen, J. Quantitative EEG for the detection of brain ischemia. *Crit. Care* **2012**, *16*, 216. [CrossRef] [PubMed]
27. Kenneth, G. Emergency EEG and Continuous EEG Monitoring in Acute Ischemic Stroke. *J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc.* **2004**, *21*, 341.
28. Burghaus, L.; Hilker, R.; Dohmen, C.; Bosche, B.; Winhuisen, L.; Galldiks, N.; Heiss, W.D. Early electroencephalography in acute ischemic stroke: Prediction of a malignant course? *Clin. Neurol. Neurosurg.* **2007**, *109*, 45–49. [CrossRef]
29. Singer, D.E.; Chang, Y.; Borowsky, L.H.; Fang, M.C.; Pomernacki, N.K.; Udaltsova, N.; Reynolds, K.; Go, A.S. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: The ATRIA study stroke risk score. *J. Am. Heart Assoc.* **2013**, *2*, e000250. [CrossRef]
30. Gage, B.F.; Waterman, A.D.; Shannon, W.; Boechler, M.; Rich, M.W.; Radford, M.J. Validation of clinical classification schemes for predicting stroke: Results from the National Registry of Atrial Fibrillation. *JAMA* **2001**, *285*, 2864–2870. [CrossRef] [PubMed]
31. Smith, A.; Patterson, C.; Yarnell, J.; Rumley, A.; Ben-Shlomo, Y.; Lowe, G. Which hemostatic markers add to the predictive value of conventional risk factors for coronary heart disease and ischemic stroke? The Caerphilly Study. *Circulation* **2005**, *112*, 3080–3087. [CrossRef] [PubMed]
32. Wolf, P.A.; D'Agostino, R.B.; Belanger, A.J.; Kannel, W.B. Probability of stroke: A risk profile from the Framingham Study. *Stroke* **1991**, 22, 312–318. [CrossRef]
33. Chambless, L.E.; Heiss, G.; Shahar, E.; Earp, M.J.; Toole, J. Prediction of ischemic stroke risk in the Atherosclerosis Risk in Communities Study. *Am. J. Epidemiol.* **2004**, *160*, 259–269. [CrossRef] [PubMed]
34. Chien, K.; Su, T.; Hsu, H.; Chang, W.; Chen, P.; Sung, F.; Chen, M.; Lee, Y. Constructing the prediction model for the risk of stroke in a Chinese population: Report from a cohort study in Taiwan. *Stroke* **2010**, *41*, 1858–1864. [CrossRef] [PubMed]
35. Jee, S.H.; Park, J.W.; Lee, S.Y.; Nam, B.H.; Ryu, H.G.; Kim, S.Y.; Yun, J.E. Stroke risk prediction model: A risk profile from the Korean study. *Atherosclerosis* **2008**, *197*, 318–325. [CrossRef] [PubMed]
36. Arslan, A.K.; Colak, C.; Sarihan, M.E. Different medical data mining approaches based prediction of ischemic stroke. *Comput. Methods Programs Biomed.* **2016**, *130*, 87–92. [CrossRef]
37. Goyal, M. Long Short-Term Memory Recurrent Neural Network for Stroke Prediction. In Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition, New York, NY, USA, 15–19 July 2018.
38. Xie, J.; Wang, Z.; Yu, Z. Enabling Efficient Stroke Prediction by Exploring Sleep Related Features. In Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 8–12 October 2018; pp. 452–461.
39. Faught, E. Current role of electroencephalography in cerebral ischemia. *Stroke* **1993**, *24*, 609. [CrossRef]
40. Claassen, J.; Hirsch, L.J.; Kreiter, K.T.; Du, E.Y.; Connolly, E.S.; Emerson, R.G.; Mayer, S.A. Quantitative continuous EEG for detecting delayed cerebral ischemia in patients with poor-grade subarachnoid hemorrhage. *Dkgest World Latest Med. Inf.* **2005**, *115*, 2699–2710. [CrossRef]
41. Macdonell, R.A.; Donnan, G.A.; Bladin, P.F.; Berkovic, S.F.; Wriedt, C.H.R. The Electroencephalogram and Acute Ischemic Stroke: Distinguishing Cortical From Lacunar Infarction. *Arch. Neurol.* **1988**, *45*, 520. [CrossRef]

42. Lansberg, M.G.; Thijs, V.N.; O'Brien, M.W.; Ali, J.O.; de Crespigny, A.J.; Tong, D.C.; Moseley, M.E.; Albers, G.W. Evolution of apparent diffusion coefficient, diffusion-weighted, and T2-weighted signal intensity of acute stroke. *Ajnr Am. J. Neuroradiol.* **2001**, *22*, 637–644.
43. Rathakrishnan, R.; Gotman, J.; Dubeau, F.; Angle, M. Using Continuous Electroencephalography in the Management of Delayed Cerebral Ischemia Following Subarachnoid Hemorrhage. *Neurocritical Care* **2011**, 14, 152–161. [CrossRef]
44. Molnár, M.; Csuhaj, R.; Horváth, S.; Vastagh, I.; Gaál, Z.A.; Czigler, B.; Nagy, Z. Spectral and complexity features of the EEG changed by visual input in a case of subcortical stroke compared to healthy controls. *Clin. Neurophysiol.* **2006**, *117*, 771–780. [CrossRef] [PubMed]
45. Lederman, R.J. Bradley's neurology in clinical practice. *JAMA* **2012**, *308*, 1694. [CrossRef]
46. Zhang, L.; He, C. *Quantitative Methods for Detecting Cerebral Infarction from Multiple Channel EEG Recordings*; Springer: Berlin/Heidelberg, Germany, 2012.
47. Wang, Z.; Guo, B.; Yu, Z.; Zhou, X. Wi-Fi CSI-Based Behavior Recognition: From Signals and Actions to Activities. *IEEE Commun. Mag.* **2018**, *56*, 109–115. [CrossRef]
48. Motamedi-Fakhr, S.; Moshrefi-Torbati, M.; Hill, M.; Hill, C.M.; White, P.R. Signal processing techniques applied to human sleep EEG signals—A review. *Biomed. Signal Process. Control* **2014**, *10*, 21–33. [CrossRef]
49. Wolpert, E.A. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. *Arch. Gen. Psychiatry* **1969**, *20*, 246–247. [CrossRef]
50. Hartmann, E. The 90-Minute Sleep-Dream Cycle. *Arch. Gen. Psychiatry* **1968**, *18*, 280. [CrossRef] [PubMed]
51. Available online: https://www.howsleepworks.com (accessed on 10 January 2020).
52. Kim, B.J.; Lee, S.H.; Shin, C.W.; Ryu, W.S.; Kim, C.K.; Yoon, B.W. Ischemic Stroke During Sleep Its Association With Worse Early Functional Outcome. *Stroke* **2011**, *42*, 1901. [CrossRef]
53. van Putten, M.J.; Hofmeijer, J. EEG Monitoring in Cerebral Ischemia: Basic Concepts and Clinical Applications. *J. Clin. Neurophysiol.* **2016**, *33*, 203. [CrossRef]
54. Palma, J.A.; Urrestarazu, E.; Iriarte, J. Sleep loss as risk factor for neurologic disorders: A review. *Sleep Med.* **2013**, *14*, 229–236. [CrossRef] [PubMed]
55. Brown, D.L.; Feskanich, D.; Sánchez, B.N.; Rexrode, K.M.; Schernhammer, E.S.; Lisabeth, L.D. Rotating night shift work and the risk of ischemic stroke. *Am. J. Epidemiol.* **2009**, *169*, 1370–1377. [CrossRef]
56. Pasic, Z.; Smajlovic, D.; Dostovic, Z.; Kojic, B.; Selmanovic, S. Incidence and Types of Sleep Disorders in patients with Stroke. *Med. Arh.* **2011**, *65*, 225–227. [CrossRef] [PubMed]
57. Chen, Y.K.; Lu, J.Y.; Mok, V.C.; Ungvari, G.S.; Chu, W.C.; Wong, K.S.; Tang, W.K. Clinical and radiologic correlates of insomnia symptoms in ischemic stroke patients. *Int. J. Geriatr. Psychiatry* **2011**, *147*, S21–S22. [CrossRef] [PubMed]
58. Lee, K. Sleep-Wake Patterns during the Acute Phase after First-Ever Stroke. *Stroke Res. Treat.* **2011**, *2011*, 936298.
59. Giubilei, F.; Iannilli, M.; Vitale, A.; Pierallini, A.; Sacchetti, M.L.; Antonini, G.; Fieschi, C. Sleep patterns in acute ischemic stroke. *Acta Neurol. Scand.* **2010**, *86*, 567–571. [CrossRef]
60. Hao, L.; Fei, W.; Ximing, L. The macrostructure of sleep in patients with stroke. *Jiangxi Med. J.* **2008**, *10*.
61. Available online: http://www.scholarpedia.org/article/Electroencephalogram (accessed on 10 January 2020).
62. Šušmáková, K.; Krakovská, A. Discrimination ability of individual measures used in sleep stages classification. *Artif. Intell. Med.* **2008**, *44*, 261–277. [CrossRef]
63. Available online: https://imotions.com/blog/eeg (accessed on 10 January 2020).
64. Wijaya, S.K.; Badri, C.; Misbach, J. Electroencephalography (EEG) for detecting acute ischemic stroke. In Proceedings of the 2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME), Bandung, Indonesia, 2–3 November 2016.
65. Al-Qazzaz, N.K.; Ali, S.H.B.M.; Ahmad, S.A.; Islam, M.S.; Escudero, J. Discrimination of stroke-related mild cognitive impairment and vascular dementia using EEG signal analysis. *Med. Biol. Eng. Comput.* **2017**, *56*, 137–157. [CrossRef] [PubMed]
66. Agius Anastasi, A.; Falzon, O.; Camilleri, K.; Vella, M.; Muscat, R. Brain Symmetry Index in Healthy and Stroke Patients for Assessment and Prognosis. *Stroke Res. Treat.* **2017**, *2017*, 8276136. [CrossRef] [PubMed]
67. van Putten, M.J.; Tavy, D.L. Continuous quantitative EEG monitoring in hemispheric stroke patients using the brain symmetry index. *Dkgest World Latest Med. Informat.* **2005**, *35*, 2489–2492. [CrossRef] [PubMed]
68. Liu, S.; Guo, J.; Meng, J.; Wang, Z.; Yao, Y.; Yang, J.; Ming, D. Abnormal EEG complexity and functional connectivity of brain in patients with acute thalamic ischemic stroke. *Comput. Math. Methods Med.* **2016**. [CrossRef] [PubMed]
69. Guo, C.; Lu, F.; Liu, S.; Xu, W. Sleep EEG Staging Based on Hilbert-Huang Transform and Sample Entropy. In Proceedings of the 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 12–14 December 2015; pp. 442–445.
70. Liu, F.; Zhou, X.; Wang, Z.; Ni, H.; Wang, T. OSA-weigher: An automated computational framework for identifying obstructive sleep apnea based on event phase segmentation. *J. Ambient Intell. Humaniz. Comput.* **2018**, *10*, 1937–1954. [CrossRef]
71. Molina-Picó, A.; Cuesta-Frau, D.; Aboy, M.; Crespo, C.; Miró-Martínez, P.; Oltra-Crespo, S. Comparative study of approximate entropy and sample entropy robustness to spikes. *Artif. Intell. Med.* **2011**, *53*, 97–106. [CrossRef]
72. Adhi, H.A.; Wijaya, S.K.; Badri, C.; Rezal, M. Automatic detection of ischemic stroke based on scaling exponent electroencephalogram using extreme learning machine. *J. Phys. Conf. Ser.* **2017**, *820*, 012005. [CrossRef]

73. Hwa, R.C.; Ferree, T.C. Stroke detection based on the scaling properties of human EEG. *Phys. A Stat. Mech. Appl.* **2004**, *338*, 246–254. [CrossRef]
74. Lempel, A.; Ziv, J. On the complexity of finite sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81. [CrossRef]
75. Zhang, Y.; Wang, C.; Sun, C.; Zhang, X.; Wang, Y.; Qi, H.; Ming, D. Neural complexity in patients with poststroke depression: A resting EEG study. *J. Affect. Disord.* **2015**, *188*, 310–318. [CrossRef]
76. Bai, Y.; Liang, Z.; Li, X. A permutation Lempel-Ziv complexity measure for EEG analysis. *Biomed. Signal Process. Control* **2015**, *19*, 102–114. [CrossRef]
77. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Volume 14, pp. 1137–1145.
78. Mohamed, H.; Mabrouk, M.S.; Sharawy, A. Computer aided detection system for micro calcifications in digital mammograms. *Comput. Methods Programs Biomed.* **2014**, *116*, 226–235. [CrossRef] [PubMed]
79. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
80. Meng, Z.; Xu, Y.; Zheng, Y.; Zhu, Y.; Jia, Y.; Chen, S. Inversion of lunar regolith layer thickness with CELMS data using BPNN method. *Planet. Space Sci.* **2014**, *101*, 1–11. [CrossRef]
81. Heo, J.; Yoon, J.G.; Park, H.; Kim, Y.D.; Nam, H.S.; Heo, J.H. Machine learning—Based model for prediction of outcomes in acute stroke. *Stroke* **2019**, *50*, 1263–1265. [CrossRef] [PubMed]
82. Available online: https://www.physionet.org (accessed on 10 January 2020).
83. Quan, S.F.; Howard, B.V.; Iber, C.; Kiley, J.P.; Nieto, F.J.; O'Connor, G.T.; Rapoport, D.M.; Redline, S.; Robbins, J. The Sleep Heart Health Study: Design, rationale, and methods. *Sleep* **1997**, *20*, 1077.
84. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [CrossRef]
85. Leevy, J.L.; Khoshgoftaar, T.M.; Bauder, R.A.; Seliya, N. A survey on addressing high-class imbalance in big data. *J. Big Data* **2018**, *5*, 42. [CrossRef]
86. Ozenne, B.; Subtil, F.; Maucort-Boulch, D. The precision—Recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* **2015**, *68*, 855–859. [CrossRef] [PubMed]
87. Wang, Y.J.; Li, Z.X.; Gu, H.Q.; Zhai, Y.; Jiang, Y.; Zhao, X.Q.; Zhao, J.Z. China Stroke Statistics 2019: A Report From the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations. *Stroke Vasc. Neurol.* **2020**, *5*. [CrossRef]
88. Sharma, S.; Culebras, A. Sleep apnoea and stroke. *Stroke Vasc. Neurol.* **2016**, *1*, 185–191. [CrossRef] [PubMed]
89. Denny, M.C.; Boehme, A.K.; Dorsey, A.M.; George, A.J.; Yeh, A.D.; Albright, K.C.; Martin-Schild, S. Wake-up strokes are similar to known-onset morning strokes in severity and outcome. *J. Neurol. Neurol. Disord.* **2014**, *1*, 1.
90. Zappasodi, F.; Olejarczyk, E.; Marzetti, L.; Assenza, G.; Pizzella, V.; Tecchio, F. Fractal dimension of EEG activity senses neuronal impairment in acute stroke. *PLoS ONE* **2014**, *9*, e100199.
91. Kannathal, N.; Acharya, U.R.; Lim, C.M.; Sadasivan, P.K. Characterization of EEG—A comparative study. *Comput. Methods Programs Biomed.* **2005**, *80*, 17–23. [CrossRef] [PubMed]