IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 June 2024, accepted 10 July 2024, date of publication 16 July 2024, date of current version 23 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3429157

APPLIED RESEARCH

Stroke Prediction Using Deep Learning and Transfer Learning Approaches

DONG-HER SHIH[1], YI-HUEI WU[1], TING-WEI WU[1], HUEI-YING CHU[1], AND MING-HUNG SHIH[2]

[1]Department of Information Management, National Yunlin University of Science and Technology, Douliu, Yunlin 64002, Taiwan
[2]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA

Corresponding author: Ting-Wei Wu (wutingw@yuntech.edu.tw)

This work was supported in part by Taiwan National Science and Technology Council under Grant NSTC 112-2410-H-224-007.

ABSTRACT Stroke is one of the leading causes of death and disability worldwide. The ideal solution to the stroke problem is to prevent it in advance by controlling metabolic factors, atrial fibrillation, hypertension, smoking, Etc. However, unless the physiological indicators are abnormal, it is difficult for medical personnel to decide whether special precautions are necessary for a patient based solely on monitoring the potential patient. There was a great category imbalance between stroke and non-stroke patients, so this study tried to use various techniques to solve the problem of categorical unbalanced stroke prediction problem. Then, deep learning models were used to predict whether the patients would have a stroke. Finally, the classification experiment is carried out through transfer learning to observe whether the evaluation metrics are further improved. According to the experimental results, this study effectively reduced the false negative rate (FNR) and false positive rate (FPR) of stroke prediction and improved the overall accuracy of stroke prediction through the category imbalance treatment and deep learning method.

INDEX TERMS Machine learning, deep learning, transfer learning, stroke prediction.

I. INTRODUCTION

According to the World Stroke Organization, 12.2 million people will have a stroke this year, and about 6.5 million will die from it, making it the leading cause of disability and one of the world's leading causes of death. Stroke deprives people of their lives, jobs, income, and social connections. In Europe, the cost of lost productivity after stroke was EUR 12 billion in 2017, and healthcare costs were estimated at EUR 27 billion [1]. About 3 to 4 percent of healthcare spending in Western world goes on stroke. In the United States, the average lifetime cost per capita of ischemic stroke, including inpatient care, rehabilitation, and follow-up care, is estimated at $140,048 [2], [3]. Stroke is the second leading cause of death internationally, and stroke can significantly affect the body's functioning, especially memory, mobility, visual perception, vocalization, and cognition, which is why it has a severe impact on all aspects of life.

A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or bursts (or ruptures). Stroke is the leading cause of death in the United States, killing approximately 160,000 Americans yearly, or one in three casualties [4]. The incidence of stroke increases with age. Compared with patients with other chronic diseases, stroke patients have more extended hospital stays, higher readmission rates, and higher medical costs [5], [6]. According to the American Heart Association, stroke is emerging as a serious health problem due to its exceptionally high mortality rate. [7]. In addition, the cost of hospitalization for stroke continues to increase [8]. As a result, there is an in-creased need for stroke prediction with advanced technologies that can assist in clinical diagnosis, treatment, prediction of clinical events, recommendation of promising therapeutic interventions, rehabilitation programs, etc. [9].

Artificial intelligence (AI) is widely used in the medical field. Its real-time and accurate characteristics can reduce risks and improve efficiency. It can be combined with AI in image recognition, operation, drug and therapy development,

The associate editor coordinating the review of this manuscript and approving it for publication was Muammar Muhammad Kabir.

etc. With the coming of an aging society, AI-assisted medical treatment can significantly improve the quality of medical treatment, reduce the burden of medical personnel and assist doctors in diagnosis and treatment. In the past decade, more studies have investigated the application of Machine Learning (ML) models in stroke prediction. Liu et al. [10] studied unbalanced data and proposed a hybrid machine learning (ML) method to predict stroke; they achieved a 71.6% accuracy and a 19.1% false-negative rate using physiological data. In addition, Liang et al. [1] tested six ML classifiers using an interpretable machine learning pipeline based on the same data and finally selected a Multilayer perceptron (MLP) classifier as the best ML model. Dritsas and Trigka [11] use ML techniques to forecast the possibility of stroke and suggested an approach for creating an efficient binary classification ML model for stroke incidence. The majority voting and stacking methods are applied to achieve the accuracy of machine learning classification. Cheon et al. [12] compared five ML methods and found a good result of prediction. It can be helpful for patients and physicians to prescreen for possible strokes. As stroke cases rise, society will face a growing economic burden. However, prompt treatment can significantly improve stroke outcomes. Awareness of stroke warning signs and taking appropriate action during a stroke event are essential to improve outcomes [12].

Most of the existing research in the prediction field involves complete and class-balanced datasets, but only some medical datasets can meet such requirements. In actual clinical practice, stroke datasets are inherently class-imbalanced [10]. The stroke dataset used in this study has 43,400 instances, and there are only more than 700 stroke patients. Such category imbalance is an important problem for data analysis. A category imbalance dataset refers to a dataset in which the number of one or some classes is more than the number of other classes, and the distribution of samples among the classes is unbalanced. Among them, the rare class is called the minority class, while the universal class is called the majority class [13]. Such cases usually occur in several fields, including fraud detection, text classification, medical diagnosis, and many others [14]. Generally, any domain involving rare events will produce a category imbalance. In healthcare, for example, diagnosing patients with a disease is more important, even if they represent only a tiny percentage of all patients. In fraud detection, although only a small percentage of transactions are fraudulent, they cause the most losses. Therefore, it is often more necessary to correctly classify minority class instances than majority class instances [15]. The most common way to deal with category imbalance data is through sampling. Sampling can be divided into two categories: under-sampling and over-sampling. In under-sampling, sampling random portions of the majority classes to balance the minority classes can be implemented in various ways. In over-sampling, however, the size of a few classes increases to the level of a majority class [15].

Deep learning (DL) methods have gradually replaced ML in recent years, so this study used the DL framework to conduct subsequent stroke prediction research. This study adopted the Kaggle stroke dataset with an enormous number of cases. Transfer learning is a technique in machine learning that involves using a model trained on one task as a foundation for a different but related task. Rather than starting from scratch, transfer learning utilizes the knowledge obtained from solving one problem to tackle a different yet related problem. Firstly, the EM algorithm was used to fill in the missing values of the stroke dataset. Random under-sampling was used to deal with the category imbalance of the stroke data set, and finally, a deep learning framework was used to predict stroke, then transfer learning is carried out to observe whether the evaluation metrics have improved. In this study, the random under-sampling method is adopted to deal with category imbalance, which considers the false positive rate (FPR) and false negative rate (FNR), because they play an essential role in medical applications [16]. FPR can result in expensive additional tests and unnecessary medical treatment, while FNR may put patients in danger of not receiving timely and appropriate treatment due to undiagnosed illness. This is why it is crucial to minimize both FPR and FNR [1]. Therefore, how to reduce the FPR and the FNR is the most significant research focus of this study. Understanding the mechanisms behind stroke occurrence through the advanced technology of this study is a valuable tool for the early detection of hidden threats and preventive measures to improve patients' quality of life and reduce the risk of stroke. The motivation of this study is to improve the performance of the evaluation metrics by combining deep learning and transfer learning with undersampling technology through clinically unbalanced datasets. In this study, the second section is the preliminary of stroke and deep learning introduction. The third section covers the Materials and Methods, the fourth section presents the results and discussions, and the final section is the conclusion.

II. PRELIMINARY

A. RELATED WORK ON STROKE PREDICTION

More and more studies have investigated the application of machine learning models in stroke prediction. The relevant research on stroke prediction is shown in TABLE 1. Alanazi et al. [17] used three different data sampling methods to develop and predict the model of the National Health and nutrition examination survey dataset, and the accuracy of the random forest prediction method of data resampling reached 96%. Liang et al. [1] suggest an easy-to-interpretable machine learning pipeline for predicting stroke using clinical data. Biswas et al. [18] developed a hybrid stroke prediction application based on ML, preprocessed the dataset to remove missing values and outliers, and built web and mobile applications.

Sailasya and Kumari [19] used Kaggle's stroke dataset to successfully predict stroke performance across a variety

TABLE 1. Research on stroke prediction.

Dataset	Data preprocessing	Results	Authors
NCHS	Data interpolation, data resampling	RF : 96% DT : 93%	[17]
Cerebral Stroke Prediction-imbalance dataset	Remove outliers and delete missing values	LR : 73.52% XGBoost : 72.58% MLP : 70.85%	[1]
	Remove unwanted noise, missing values, outliers, and tag coding	SVM : 99.99% RF : 99.87% KNN : 98.82%	[18]
Stroke Prediction Dataset	Data interpolation (average)	KNN: 80% SVM : 80% Naïve Bayes : 82%	[19]
	Missing values are processed, minimum-maximized, and scaled	Nearest neighbor: 99.61%	[4]
Cerebral Stroke Prediction-imbalance dataset	Through RFR interpolation, PCA-K-means preprocessing	FNR : 19.1 ±1.7 Specificity : 32.6 G-mean : 46.9 Accuracy : 71.6 ± 1.2	[10]

of physiological attributes using various machine learning methods after processing missing values. In another study, Hassan et al. [4] used three different models after data preprocessing of the same dataset. The accuracy of the proposed model can reach 99.61% when using the nearest neighbor method. Liu et al. [10] proposed a mixed ML method to predict stroke based on unbalanced data. Physiological data (783 stroke patients among 43.400 subjects) were used for training, and random forest regression was used to estimate the missing value before classification. By using the deep neural network (DNN) optimized by AutoHPO, an accuracy rate of 71.6% was achieved. However, the accuracy of their study of stroke prediction still has room for improvement. TABLE 1 contains different datasets, so its accuracy will be different. The most famous data set is the Cerebral Stroke Prediction-imbalance dataset from Kaggle, and the subsequent development of this study is based on this dataset.

B. CATEGORY IMBALANCE

Over the past decade, a variety of algorithms have been proposed to solve the problem of unbalanced data classification. When data is highly unbalanced, it means that one class has significantly fewer observations than the other classes, resulting in extremely skewed distributions. Unbalanced datasets often appear in several research areas, including fraud detection, text classification, medical diagnosis, and many others [13]. In the healthcare field, it is more important to diagnose patients with certain diseases, even if they represent only a small percentage of all patients [15]. In ML and DL studies, un-der-sampling and oversampling methods are often used to adjust class distribution. [20].

Under-sampling method aims to balance data by conducting random sampling from data with multiple classes and selecting representative instances from most classes to achieve the same amount of data from two classes [21]. In the under-sampling technique, parts of the main classes are removed from the training data, making the size of the training dataset more similar to or comparable to a few classes [22]. The simplest but most effective method in the under-sampling algorithm is random under-sampling (RUS), which involves the random elimination of most class examples. RUS, like ROS, can be easily ex-tended to process multi-class data [21], [23]. According to Yen and Lee [24], a subset of MA (majority classes) is randomly selected and then combined with MI (minority classes) as a training set. The primary under-sampling technique for arbitrarily eliminating most class examples to balance a dataset is called random under-sampling (RUS) [25]. Guo et al. [26] reveal that this method is the processing of beneficial information, which may prove to be crucial in the later classification stage. Other under-sampling methods are the Condensed Nearest Neighbor Rule, Wilson edited the nearest neighbor rule, neighborhood cleaning rule, Tomek links, and One-Sided Selection (OSS) [27], [28], [29]. The biggest difference with over-sampling is that under-sampling will not create composite data, and most un-der-sampling algorithms can process mixed and incomplete data [30]. Therefore, under-sampling technology is adopted in this study for stroke prediction.

C. OVER-SAMPLING AND UNDER-SAMPLING

In sample processing, increasing the number of a few categories is called over-sampling, whereas reducing the number of significant categories is called under-sampling. In machine learning, oversampling and under-sampling are commonly used to deal with data sets with unbalanced category distribution [31]. In the over-sampling method, although unnecessary information loss is avoided by increasing the number of a few categories, it also has the opportunity to increase the noise and outlier together [31]. In addition, if it is applied to a large-scale dataset, with the increase in training data, technical difficulties will be increased, including increased training time, the need for sufficient memory storage, etc. [32]. Another common resampling strategy is un-der-sampling, which aims to alter the distribution of categories in the training data. This is a separate data preprocessing step that can directly balance the data prior to training the classifier. The total number of training data can be significantly reduced, and training costs can be reduced. However, experience has proved it is a very effective sampling method [32], [33]. Luengo et al. [33] processed many unbalanced datasets with over-sampling and under-sampling methods and then performed classification performance by the famous classification methods C4.5 and PART and found that the dataset treated with under-sampling could get better results. Zughrat et al. [32] used the iterative SVM method to classify rail data with Category imbalance, and the results showed that the under-sampling method could suppress the number of support vectors and increase the performance of SVM. Therefore, this study chose the under-sampling method to process the class imbalance data of stroke. Furthermore, compare different performance evaluation metrics.

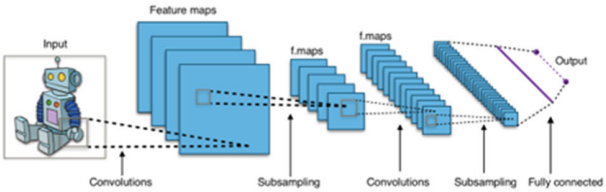

FIGURE 1. Convolutional neural network (CNN).

D. DEEP LEARNING AND CONVOLUTIONAL NEURAL NETWORKS

In recent years, DL technology has become very famous and has been applied in various fields [34], [35]. In the medical field, researchers are utilizing deep learning techniques for various tasks, including feature extraction, classification, image segmentation, disease prognosis prediction, and overall survival prediction due to the success of deep learning [36].

1) CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional Neural network (CNN) is an effective deep learning technique [37]. In medical image analysis research, CNN is commonly utilized due to its ability to preserve spatial relationships while filtering input images [38]. CNN uses a feedforward neural network to identify image features. The units in each layer are fully connected to the adjacent layer. Only specific units are connected to the fully connected layer, which is responsible for classifying image features [39].

The overall architecture of CNN is shown in Figure 1. CNN has multiple layers, which are mainly divided into the convolutional layer, pooling layer, and fully connected layer [40]. CNN has many advantages: First, local connection, each neuron no longer connects all the neurons in the upper layer but only connects a few neurons, effectively reducing parameters and speeding up the convergence speed. Another technique is weight sharing, where connections can have the same weight, resulting in fewer parameters. Additionally, to reduce the image's size, the pooling layer uses the principle of local correlation to down-sample the image while retaining valuable information and minimizing data. It can also reduce the number of parameters by removing trivial functions. These three attractive features make CNN one of the most representative algorithms in the deep learning field [41]. The definition equation (1), (2) of convolutional neural networks are as follows:

$$f(x) = \max(0, x) \tag{1}$$

$$f_{X,Y}(S) = \max_{a,b=0} S_{2X+a,2Y+b} \tag{2}$$

where f(x) is defined as a linear rectifier layer and $f_{x,y}(S)$ Is the pooling layer.

2) ARTIFICIAL NEURAL NETWORK (ANN)

Now established as the field of computer science, artificial intelligence (AI) is dedicated to producing application software that can perform complex, intelligent calculations

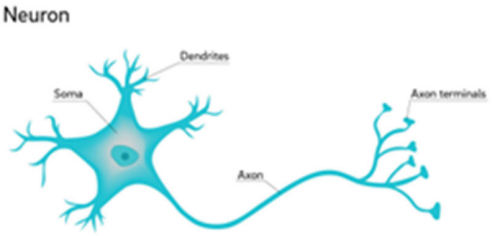

FIGURE 2. Brain neuron.

similar to those performed by the human brain on a daily basis. Artificial neural networks are similar to digital models of the human brain and can simulate the way the brain processes information. Ai learns (or trains) through appropriate learning examples, just like humans, not through programming [42]. Artificial neural networks are made up of processing units called neurons. Artificial neurons attempt to replicate the structure and behavior of natural neurons. Neurons consist of inputs (dendrites) and outputs (synapses via axons). Neurons have a function that determines neuron activation [43], The Brain neuron construction is shown in Figure 2.

Artificial neural networks are commonly utilized for complex tasks such as nonlinear function mapping, image processing and recognition, pattern recognition, and classification. Feedforward network is a common neural network. The feedforward network includes an input layer in which the input to the problem is received. The hidden layer, where the relationship between input and output is determined and represented by synaptic weights, and the output layer, performs the output of the problem [44].

3) DEEP NEURAL NETWORK (DNN)

A DNN architecture is a collection of neurons organized in a sequence of multiple layers, where neurons receive neuronal activation from the previous layer as input and perform various calculations (e.g., weighted and followed by nonlinear activation of inputs). The neurons of the network realize complex nonlinear mapping from input to output. This mapping is learned from large amounts of data by adapting the weight of each neuron using a technique called error back propagation [45], The main architecture of DNN is shown in Figure 3.

Feedforward DNNs are trained to automatically learn transformations that map inputs to outputs. Each hidden layer is a learning feature, which helps with the discrimination task. They consist of an input layer that receives data through an input vector and two or more hidden layers. These hidden layers transform the previous layer's output, resulting in a higher level of representation for the input layer. The output layer calculates the DNN output [46].

III. MATERIALS AND METHODS

A. DATASET

The dataset of this study was taken from Kaggle's public dataset Cerebral Stroke Prediction-Imbalanced

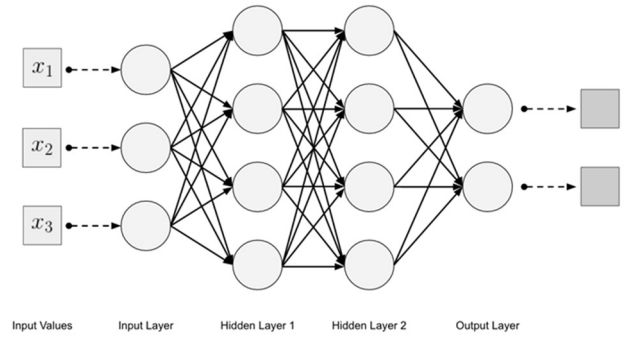

FIGURE 3. Major architectures of DNN.

TABLE 2. Description of the input variables.

Variables	Names	Definition	Data type
x_1	Age	Age of patient	continuous
x_2	BMI	Body mass index (BMI) is a measure of body fat	continuous
x_3	Gender	Female/male	category
x_4	Average glucose	Average blood sugar is an estimated average of blood sugar	continuous
x_5	Work	Never worked/child/government/self-employed/private	category
x_6	Residence	Rural/urban	category
x_7	Smoking status	Never smoked/ever smoked/smoker/unknown	category
x_8	Heart disease	No/Yes	category
x_9	Married	No/Yes	category
x_{10}	Hypertension	No/Yes	category

Datase [47] (https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset, accessed on October 12, 2022). The dataset was typically a category imbalance, containing 10 risk factors for stroke, as well as the target variable ''stroke''. Of the 43,400 subjects, 783 patients experienced a stroke, and the remaining 42,617 patients had no symptoms of a stroke. The dataset is incomplete, and 30% smoking status items and 3% body mass index (BMI) items are missing. Because the datasets come from the medical field, it is not uncommon for there to be a sizable category imbalance in the number of people with and without the disease. The detailed data set description of variables in this study is shown in TABLE 2. In this study, we utilized the RandomUnderSampler function from Python's imbalanced-learn package to undersample the data. We performed data preprocessing and random under sampling to create the experimental dataset, which comprised a total of 1276 data sets with a stroke to nonstroke ratio of 1:1.

This study also adopts transfer learning to observe whether the evaluation index can be improved. Transfer learning requires pre-training a model from the source dataset and then learning the model's weight to the target dataset for training. The transfer learning Dataset for this study was taken from Kaggle's public dataset ''Dataset Surgical binary. classification'' (https://www.kaggle.com/datasets/omnamahshivai/surgical-dataset-binary-classification/code?select=Surgica l-deepnet.csv, accessed May 07, 2023). The dataset, which was used as a pre-training for transfer learning in this study, included 24 risk factors for morbidity and the target variable ''complications.'' Of the 14,635 participants, 3690 patients experienced complications, and the remaining 10,945 patients had no symptom-related complications. Detailed data set descriptions are shown in TABLE 3.

TABLE 3. Description of the transfer learning dataset.

Variables	Names	Data type
tx_1	ahrq_ccs	continuous
tx_2	age	continuous
tx_3	gender	category
tx_4	race	category
tx_5	asa_status	category
tx_6	bmi	continuous
tx_7	baseline_cancer	category
tx_8	baseline_cvd	category
tx_9	baseline_dementia	category
tx_{10}	baseline_diabetes	category
tx_{11}	baseline_digestive	category
tx_{12}	baseline_osteoart	category
tx_{13}	baseline_psych	category
tx_{14}	baseline_pulmonary	category
tx_{15}	baseline_charlson	continuous
tx_{16}	mortality_rsi	continuous
tx_{17}	complication_rsi	continuous
tx_{18}	ccsMort30Rate	continuous
tx_{19}	ccsComplicationRate	continuous
tx_{20}	hour	continuous
tx_{21}	dow	category
tx_{22}	month	category
tx_{23}	moonphase	category
tx_{24}	mort30	category
Ty	complication	category

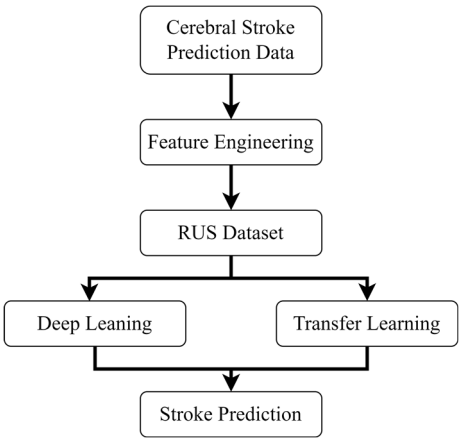

FIGURE 4. Main architectures.

B. RESEARCH PROCESS

The main architecture of the research process in this study is illustrated in Figure 4. The first step involves feature engineering, detailed in next section, followed by the application of deep learning models and transfer learning models for stroke prediction.

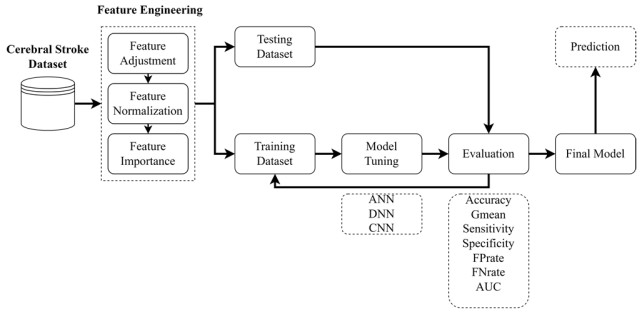

FIGURE 5. Method I: Deep learning process.

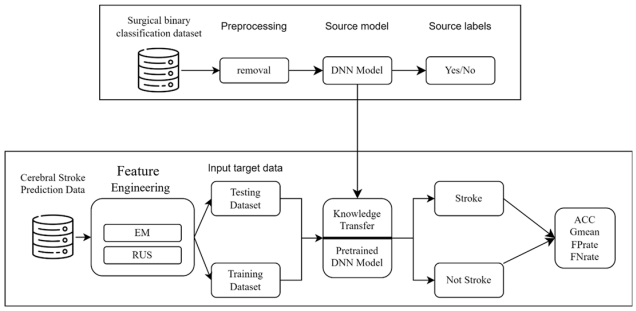

FIGURE 6. Method II: Transfer learning process.

The flowchart of Deep learning process, named as Method I, is shown in Figure 5. The first step is feature engineering. The feature adjustment method in feature engineering processes the missing values of the data set, and then the min-max data is normalized through the feature normalization step in feature engineering. Then, feature selection methods from WEKA3.8 are used to determine the feature importance. Secondly, under-sampling is applied to process the category imbalance of the dataset. The under-sampling method will screen representative data from most classes of data and form the dataset after the category imbalance processing. This study uses the RandomUnderSampler function in python's imbalanced-learn package for under-sampling. The data set was divided into the training set and the test set to 8:2 and input into three deep learning models (CNN, ANN, DNN). Then the parameters were adjusted to try to achieve the best evaluation results. The hyperparameters of the deep learning model adopted are presented in the results section.

Finally, the test set is fed into the deep learning architecture after training, and the prediction results of deep learning are evaluated through the evaluation metrics. The evaluation metrics are Accuracy, G-mean, FNR (false-negative rate), and FPR (false-positive rate), which will show in section III-D.

Another flowchart of transfer learning process, named as Method II, is shown in Figure 6. After the data preprocessing of the transferred learning dataset, the deep learning model is pretrained, and the parameters are adjusted to obtain the results. After the model's weight is accessed, the model is transferred to the stroke data set for model training and testing. Again, feature engineering is applied to the stoke dataset by using an EM (Expectation-on-Maximization algorithm) and RUS (Random Under Sampling) which will detailed in next section. Finally, the best evaluation results are generated after fine-tuning to observe whether there is a possibility of improvement of the stroke prediction.

C. FEATURE ENGINEERING

Feature engineering is the process of using domain knowledge to select, modify, or build features (input variables) to make machine learning algorithms work more efficiently [48]. This is a key step in the data preprocessing phase of machine learning, which converts raw data into meaningful features that improve the performance of predictive models. Possible methods of feature engineering include:

1) FEATURE SELECTION

Identify and select the most relevant features from the original data. The feature selection for this study is shown in TABLE 2.

2) FEATURE EXTRACTION

involves mathematical transformation, aggregation or application of specific algorithms to extract useful information. The input variables in TABLE 2 come from the database that has completed feature extraction.

3) FEATURE TRANSFORMATION

Modify features to better suit the needs of machine learning algorithms. These include standardizing or scaling numeric features, encoding categorical features, and handling missing values. The feature transformation used in this study is as follows:

1. Missing values: One of the most common ways to deal with missing values for incomplete data is an interpolation, replacing missing data with actual values [49]. The traditional approach is "exclusion by list" or Complete Case Analysis (CCA). CCA removes the entire case containing any missing data from the analysis [50]. This study dataset contains some obvious outliers and noise, such as age and BMI features, while an ID is an obvious redundant feature that is simply removed. Dempster et al. [51] put forward an expectation-on-Maximization algorithm (EM) for missing data interpolation. EM consists of a two-step process; the first step is called the E-step. It is the process of estimating the probability distribution of missing data completion given the model. Step M is the second step, determining the parameter estimates to maximize the log-likelihood of the complete data obtained from step E. The M step will reach convergence or reach the number of iterations as the stop criteria. Optimize the maximum likelihood of the data by repeating two steps up to coverage; estimate a value using other variables

(the expected step), then check if it is the most likely value (the maximization step). There is a strategy for filling in missing data with the EM algorithm. This algorithm calculates the mean and covariance matrix using the available data before filling in the missing values for continuous data. It repeats this process until the mean and covariance matrices no longer change significantly between iterations [52]. EM algorithms are most effective when data loss occurs randomly. Therefore, the EM algorithm can only be used to obtain more reliable estimates of missing data from highly correlated data [53].

2. Feature normalization: mathematical calculations of machine learning algorithms often require the normalization of numerical data so that the eigenvalues of high numerical numbers do not overwhelm the low ones. Min-maximum scaling is one of the most commonly used normalization techniques, where values scale between two given numbers. The most common way to do this is to scale between 0 and 1.

3. Feature importance is a critical concept in machine learning that measures how much each feature influences a model's predictions. Understanding the importance of features is helpful for model interpretation, feature selection, and model tuning. In this study, many feature selection methods of WEKA3.8 are used to determine the importance of features.

4. RUS (Random Under Sampling): It involves randomly reducing the number of instances in the majority class to match the number of instances in the minority class. This helps create a more balanced dataset, which can improve the performance of certain machine learning algorithms that are sensitive to class distribution. This study uses the RandomUnderSampler function in python's imbalanced-learn package for under-sampling.

4) DIMENSIONALITY REDUCTION

Reduce the number of features while retaining as much information as possible. Due to the high correlation of the input variables in TABLE 2, this study did not perform dimensionality reduction.

More effective feature engineering can significantly improve the performance of machine learning models by enhancing their ability to learn from data.

D. EVALUATION METRICS

In order to solve the category imbalance problem, the four measures based on the prediction results will form a confusion matrix, which can be divided into True Positive (TF), True Negative (TN), False Positive (FP), and False Negative (FN). The efficiency obtained after classification was evaluated, then the classifier was objectively evaluated by predictive Accuracy and G-Mean. In the prediction of stroke, priority should be given to false negative rate (FNR) and false positive rate (FPR) [10], [54]. The evaluation indicators are shown as follows:

Accuracy (ACC) is defined as the ratio of the number of correctly classified samples to the total number of samples in the data set. TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative. Accuracy is the ratio of correctly classified samples to the total number of samples in the evaluated dataset. This metric is one of the most commonly used for ML in medical applications but is also known for being misleading in the case of different class proportions, as simply assigning all samples to the universal category is an easy way to achieve high accuracy. The accuracy is limited to [0,1]. [55]. The accuracy equation is as follows: (3)

$$\textbf{Accuracy} = \left(\frac{(\textbf{TP} + \textbf{TN})}{(\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN})}\right) \tag{3}$$

*G-Mean:*This function helps to select the best classification option between two categories. It has the advantage of reducing the number of incorrect classifications while increasing the number of accurate ones. The G-mean is an essential measure for evaluating sensitivity and specificity [56]. The G-mean equation is as follows: (4)

$$\text{G} - \text{mean} = \sqrt{\text{sensitivity} \times \text{sepecififfity}} \tag{4}$$

FNR: The false-negative rate represents the number of samples wrongly classified as negative. For example, patients who had a stroke were incorrectly predicted not to have one. The FNR equation is as follows: (5)

$$\textbf{FNR} = \frac{\textbf{FN}}{\textbf{TP} + \textbf{FN}} \tag{5}$$

*FPR:*The false positive rate represents the number of samples wrongly classified as positive. For example, patients who had not had a stroke were incorrectly predicted to have one. The FPR equation is as follows: (6)

$$\textbf{FPR} = \frac{\textbf{FP}}{\textbf{TN} + \textbf{FP}} \tag{6}$$

IV. RESULTS AND DISCUSSION

A. RESULTS

In this study, three different deep learning models (CNN, ANN, DNN) were first adopted for classification experiments, selecting hyperparameters is a crucial step in creating effective deep-learning solutions. DL algorithms often include specific hyperparameters that control various factors, such as memory usage and execution costs. A hyperparameter is a variable that is set before a learning algorithm is applied to a context-specific dataset. The optimal number depends on the characteristics of the dataset associated with each task and each situation. There are two ways to select and optimize hyperparameters: manual selection and automatic selection. Both methods have their advantages and disadvantages. Manual selection requires a deep understanding of the model, while automatic selection algorithms require high

TABLE 4. Comprehensive experimental results.

Deep Learning Models						
CNN-Testing Result (8:2)						
TPR	TNR	ACC(t-test)	Gmean	FPR	FNR	F1
0.833	0.718	0.777	0.773	0.282	0.167	0.794
ANN-Testing Result (8:2)						
0.833	0.742	0.789	0.786	0.258	0.167	0.79
DNN-Testing Result (8:2)						
0.841	0.75	0.797	0.794	0.25	0.159	0.81
DNN-Testing Result (7:3)						
0.782	0.747	0.765	0.764	0.253	0.218	0.774
DNN-Testing Result (6:4)						
0.754	0.705	0.73	0.729	0.295	0.246	0.74
CNN 5-Fold Cross Validation Result						
TPR	TNR	ACC	Gmean	FPR	FNR	F1
0.803	0.734	0.77	0.768	0.266	0.197	0.782
ANN 5-Fold Cross Validation Result						
0.811	0.742	0.777	0.776	0.258	0.189	0.79
DNN 5-Fold Cross Validation Result						
0.818	0.766	0.793	0.792	0.234	0.182	0.803
Transfer Learning Model						
TL-DNN Result						
0.886	0.718	0.805	0.812	0.230	0.144	0.824

TABLE 5. Hyperparameters for DL Technique.

Hyperparameter	CNN	ANN	DNN
Batch size	8	16	4
Activation function	Sigmoid	Sigmoid	Sigmoid
Number of epochs	10	30	60
Optimizer algorithm	Adam	Adam	Adam
Number of hidden layers	-	-	3

computational costs. The choice between the two methods involves a trade-off [57].

The pre-processed and random under-sampled data sets were respectively used to generate the experimental results with a split ratio of 8:2 between the training set and the test set, and the detailed results are shown in TABLE 4. Three deep learning model hyperparameters in this study were set as shown in TABLE 5. According to the study on language recognition conducted by DNN [46], different hidden layers in DNN architecture were adjusted to obtain the best results. In this study, experiments were conducted in the same way, FPR, FNR, Gmean, and F1 were taken as the primary evaluation metrics, and finally, experimental results of two hidden layers were taken as the best setting. Detailed adjustment results of hidden layers are shown in TABLE 5.

According to the experimental results in TABLE 4, the results of the evaluation metrics of the three DL models are close to each other, and the experimental results of ANN and DNN are better than those of CNN with the categories of unbalanced metrics such as FPR, FNR, Gmean,

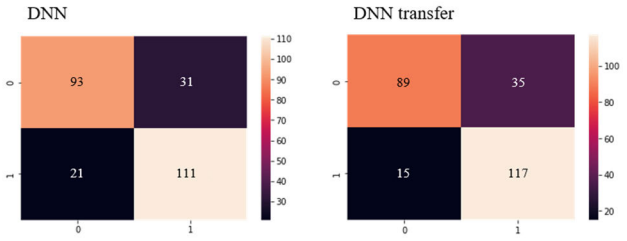

FIGURE 7. Confusion matrix of DNN and DNN transfer models.

and F1 as the criteria. Compared with the AUC curve, the AUC of ANN and DNN reached 0.8, and the evaluation result was good (https://darwin.unmc.edu/dxtests/roc3.htm, accessed on 15 January 2023). The study tested samples of different training sets using DNN models and found that those with better test results were used. The results indicated that the experiment with an 8:2 ratios between training and test sets produced the best effect. In order to evaluate whether the built model is overfitting or underfitting, according to Koehrsen [58], it can be done by establishing validation datasets or k-fold cross-validation, which allows the model to be optimized before deployment without having to use additional data. Therefore, 5-fold cross-validation was also carried out in this study, and the average value of the evaluation index was also shown in TABLE 4. Overall, the testing results of DNN model are better that the testing results of CNN and ANN.

In this study, we used Transfer Learning after conducting a deep learning model experiment to determine if the evaluation metrics could be enhanced. We pre-trained a transfer dataset and transferred the weight of the learning model to the stroke dataset. TABLE 4 shows that after applying transfer learning (TL-DNN model), the accuracy improved to 0.805, Gmean to 0.812, and FNR to 0.144. In order to compare the performance, the confusion matrix of DNN and DNN transfer models are also shown in Fig. 7, which shown that the TL-DNN transfer learning model is better than the DNN DL model. These results suggest that transfer learning can improve the performance of evaluation metrics beyond the original deep learning model. In the medical field, adopting other datasets with similar characteristics for transfer learning may help its diagnosis results. CNN related models are specifically designed to analyze images and detect patterns and spatial relationships. They are highly effective at extracting features from images with convolutional and pooling layers [38]. However, numerical data does not have the same spatial characteristics as images. For numerical data, general deep learning models are better suited to learn complex relationships in a fully connected manner. Finally, a more de-tailed assessment results of different DL models includes Multilayer Perceptron (MLP), Deep Belief Network (DBN), CNN-LSTM, and CNN-GRU are also shown for comparison in TABLE 7. And, transfer learning model (TL-DNN) shows the best result in almost metrics.

TABLE 6. Testing result with different numbers of hidden Layers.

Number of Hidden Layers	DNN-Testing Result						
	TPR	TNR	ACC	G-mean	FPR	FNR	Neuron
3	0.841	0.75	0.797	0.794	0.25	0.159	128
4	0.864	0.718	0.793	0.787	0.282	0.136	64
5	0.879	0.685	0.785	0.776	0.315	0.121	32
6	0.818	0.742	0.781	0.779	0.258	0.182	16

TABLE 7. Assessment results of different DL models.

Models	Performance Metrics						
	Sensitivity	Specificity	Accuracy	G-Mean	FPR	FNR	AUC
ANN	0.823	0.783	0.801	0.803	0.217	0.177	0.80
MLP	0.823	0.783	0.80	0.802	0.217	0.177	0.80
DBN	0.758	0.806	0.781	0.782	0.194	0.242	0.78
CNN	0.824	0.740	0.773	0.781	0.260	0.176	0.77
CNN-GRU	0.806	0.75	0.773	0.777	0.250	0.194	0.77
CNN-LSTM	0.631	0.760	0.680	0.693	0.240	0.369	0.68
DNN	0.806	0.750	0.773	0.777	0.250	0.194	0.77
TL-DNN	0.856	0.770	0.805	0.812	0.230	0.144	0.80

TABLE 8. Experimental results of outlier removal.

TPR	TNR	ACC	G-mean	FPR	FNR	F1
Condition 1 DNN (age of deletion less than 10 years and BMI greater than 60%)						
0.835	0.776	0.800	0.805	0.224	0.165	0.781
Condition 2 DNN (age less than 25 years and BMI greater than 60%)						
0.789	0.773	0.780	0.781	0.227	0.211	0.767
Condition 3 DNN (delete mean plus/minus with two standard deviations)						
0.778	0.730	0.753	0.754	0.222	0.278	0.749
Condition 4 DNN (delete mean plus/minus with three standard deviations)						
0.842	0.728	0.780	0.783	0.272	0.158	0.777

B. DISCUSSION

1) OUTLIER REMOVAL AND FEATURE IMPORTANCE

In statistics, an outlier is an observation point that is distant from other observations. The outliers can be a result of a mistake during data collection or it can be just an indication of variance in collected data. Removing outliers is a simple method to enhance the appearance and organization of data. Therefore, four different outliner conditions were adopted in this study, the removal age of condition 1 was less than ten years old, and BMI>60% [1], followed by the removal age of condition 2 was less than 25 years old and BMI>60% [59]. For Condition 3, the results of the mean value of blood glucose and BMI plus/minus two standard deviations were deleted. Finally, Condition 4, the results of the mean value of blood glucose and BMI plus/minus three standard deviations were deleted. Compared with the performance results in TABLE 8, Condition 1 is superior to Condition 2 if age and BMI are used as the main deletion conditions. The performance of Condition 4 is better than Condition 3 if the mean plus/minus standard deviation is mainly used to delete the condition. However, outlier removal in this stroke dataset didn't show a better prediction result in compare with the results of TABLE 4.

In addition, different feature selection techniques are used to illustrate the importance of features in this study. WEKA 3.8's CorrelationAttributeEval technique is used, which is based on Pearson's correlation coefficient in statistics. Another InfoGainAt-tributeEval Attribute Evaluator technique is calculating each feature's information gain. This study presents two feature selection techniques through TABLE 9 and lists the top 5 features. According to TABLE 9, the ranking results of the two feature methods are the same, however, BMI and smoking status, which are usually considered important, are not in the top 5 list.

TABLE 9. Feature importance in stroke prediction.

Feature Selection Technique	
CorrelationAttributeEval	InfoGainAttributeEval
1.Age	1.Age
2.Heart disease	2.Heart disease
3.Average glucose	3.Average glucose
4.Hypertension	4.Hypertension
5. Married	5. Married

2) CATEGORY IMBALANCE, ML, AND DL

At present, most of the medical and clinical data are in a state of category imbalance, and the patients with past illnesses are in a minority class. Therefore, it is very important to properly deal with the category imbalance data [10], [15]. In this study, random under-sampling is adopted to deal with category imbalance, and the main reason is that the oversampling may lead to bias and overly optimistic estimation of the model [60], and the evaluation index of oversampling is often focused on accuracy. Clinical needs to reduce FPR and FNR may not be met. In addition, the G-Mean evaluation metric is a measure of the balance between the classification performance of the majority and minority classes. A low G-Mean indicates the poor classification of positive cases, even if negative cases are correctly classified, which is important to avoid overfitting negative and positive cases [61]. The results of this study and previous studies are summarized in TABLE 10. Note that TABLE 10 is the study results for the same dataset (Kaggle, Cerebral Stroke Prediction Imbalanced Dataset, the original data is 43,400). RUS is used as the sampling method in this study as in [1], while AUS (Active Under Sampling) is used as the sampling method in [10]. TABLE 1 in section II uses different data sets, and their basis is different, so the accuracy rate in TABLE 1 is only for reference.

The experimental results of various ML methods (AdaBoost, RF, SVM) carried out through WEKA software are also included in TABLE 10. It can be seen from that the accuracy rate of DL research results is close to 80%, which is much higher than the 71% of previous studies with ML methods. In addition, transfer learning (TL-DNN) was also adopted in this study to apply the weights generated by other datasets to the stroke data set through pre-training and significantly improve accuracy, Gmean, and FNR. The accuracy of transfer learning approach (TL-DNN) is over 80%, which is the highest among all models. In addition, in terms of FPR and FNR, the results of transfer learning

TABLE 10. Comparison of relative studies.

Author	Data Pre-processing	Sampling	Method			
			ML	Metrics (%)	DL	Metrics (%)
[10]	RFR interpolation	AUS	Bag, RF, XGB, Avg	-	AutoHPO-based DNN	GMean : 46.9 ACC : 71.6
[1]	Delete missing value	RUS	LR, RF, XGB, KNN, SVM, MLP	-	-	GMean: 75.83 ACC : 70.85
Our Study	EM interpolation		AdaBoost	GMean: 73 ACC: 72.5	CNN	GMean: 77.3 ACC: 77.7
			RF	GMean:75.6 ACC: 75.6	ANN	GMean: 78.6 ACC: 78.9
			SVM	GMean:76.3 ACC: 77.8	DNN	GMean: 79.4 ACC: 79.7
					TL-DNN	GMean: 79.8 ACC: 80.5

approach (TL-DNN) are also the best, it effectively reduces the high FPR and FNR values of previous studies. In general, the DL approach is better than using ML approach, however, the results of transfer learning approach (TL-DNN) is the best in comparing the results of various evaluation metrics.

However, even though the effect of DL may be better compared with ML, DL requires a higher experimental environment and costs more than ML [62]. In addition, the application of a DL model may improve classification efficiency, and data preprocessing may also be one of the reasons [63]. Nevertheless, some other stroke-related studies' dataset differs from the Kaggle data set adopted in this study. For example, O 'Connell et al. [64] use ML methods to identify the gene expression pattern, which is expected to accelerate the early detection of whether patients have the possibility of stroke. Theofilatos et al. [65] identified transcription patterns of acute stroke patients by ML analysis of gene expressions. Ren et al. [66] established a diagnostic model for stroke through 9 inflammation-related genes. The NCBI data sets adopted by other studies [67], [68], [69] mostly focus on gene expression pattern recognition of stroke instead of stroke prediction, and DL techniques may be used to investigate more in the future.

C. ABLATION STUDY

Ablation study is a technique used in machine learning and deep learning to systematically remove or replace a part of a model to assess the impact of that part on the overall model performance. This approach helps to understand the components of the model and their importance to the model's predicted results. Sheikholeslami et al. [70] introduced a framework for ablation research execution. Dataset configuration C_D was first introduced. Dataset X contained n features, and dataset configuration C_D (X $\{xn\}$) represented features to be excluded from $\{x\}$. For example, C_D (X $\{x1\}$) means the data set is trained after skipping feature $x1$.

TABLE 11. Ablation study result.

Excluded Feature	Performance Metrics			
	Accuracy	Precision	Sensitivity	AUC
None (base trial)	0.783	0.778	0.778	0.840
Gender	0.775	0.750	0.77	0.842
Age	0.676	0.680	0.700	0.74
Hypertension	0.768	0.775	0.731	0.836
Heart_disease	0.775	0.774	0.760	0.840
Ever_married	0.773	0.739	0.804	0.841
Work_type	0.775	0.818	0.673	0.839
Residence_type	0.778	0.817	0.619	0.841
Avg_glucose_level	0.755	0.856	0.504	0.838
Bmi	0.783	0.817	0.645	0.841
Smoking_status	0.770	0.748	0.812	0.839

Table 11 shows the experimental results of an ablation study conducted using the DNN classification method. It can be seen that if the age variable is excluded, the accuracy of the test will be greatly reduced. Kissela et al. [71] mentioned that the population with stroke was getting younger. Stroke in younger patients may carry a greater burden of lifelong disability.

V. APPLICATION TO AN ADDITIONAL DATASET

We applied our method to an additional dataset to validate the robustness and generalizability of our approach. The additional dataset we selected for this purpose is the hospital readmissions within a 30-day period among hemodialysis patients based on monthly blood test data obtained from [72].

A. ADDITIONAL DATASET

This additional dataset gathered data from a hemodialysis unit at a Taiwan hospital spanning the years 2011 to 2022 [72], adhering to the testing protocols established by the National Kidney Foundation. The dataset excluded patients in long-term respiratory care or those not undergoing consistent long-term hemodialysis, concentrating specifically on outpatient admissions. Patients who had been treated for less than three months were excluded due to incomplete data and emergency situations. Consequently, 251 out of the initial 790 patients met the criteria for inclusion, with the objective of enhancing patient care through predictive modeling.

In total, 9367 records from these 251 hemodialysis patients were analyzed, encompassing basic demographic information, laboratory test results, and data on hospital readmissions. To ensure patient privacy, anonymized patient IDs were utilized. Monthly tests included routine blood analyses along with specific assessments for electrolytes, nutritional status, liver function, dialysis efficiency, and lipid levels. Additional comprehensive tests were conducted quarterly. The dataset

TABLE 12. Features of the additional dataset.

Feature	MEAN	STD
WBC	6.84	2.43
RBC	3.42	0.56
HbC	10.19	1.40
ALBUMIN	3.86	0.45
GOT	19.23	59.45
GPT	16.02	30.78
ALKALINE-P	93.70	52.83
CHOLESTEROL	157.69	39.96
GLUCOSE	197.87	109.7
BUN_BEFORE	67.01	19.97
BUN_AFTER	16.40	6.91
CREATININE	9.07	2.56
URIC_ACID	6.68	1.60
TOTAL_CALCIUM	9.16	0.89
P	4.99	1.59
FERRITIN	397.44	446.38
INTACT-PTH	442.41	441.21
HBA1C	7.03	2.03
KTV(Kt/V)	1.46	0.29

also examined hospital readmission data, focusing on readmissions occurring within 30 days after testing. TABLE 12 displays descriptive statistics such as mean and standard deviation for each feature used.

B. PROCESS

1) DATA PRE-PROCESSING

Given the clinical nature of the dataset, addressing missing test data and the imbalance between hospital readmissions and non-readmissions is essential. Notably, non-readmissions comprise 90.87% of the 9367 records, highlighting the need for specific techniques to manage this imbalance and the missing data effectively during model construction.

2) DATA IMBALANCE PROCESSING

Data imbalance, where one class is significantly more prevalent than another, can bias model accuracy towards the majority class. To correct this, techniques such as oversampling or undersampling are used. Research indicates that oversampling methods, particularly Synthetic Minority Over-Sampling Technique (SMOTE), are more effective than undersampling for balancing datasets.

3) HANDLING MISSING VALUES

Missing values in the dataset, often due to errors in data collection, can be addressed through deletion or interpolation. To prevent data loss, interpolation is generally preferred. This experiment employs the k-Nearest Neighbors (KNN) method for handling missing values, optimizing the K value for each model to achieve the best results.

4) CROSS-VALIDATION

K-Fold Cross-Validation is used to evaluate model generalization by dividing the dataset into K subsets. The model is trained on K-1 subsets and validated on the remaining subset, rotating this process until each subset has served as the validation set. This method provides a robust assessment of the model's performance. In this study, K is set to 10, balancing concerns about data bias and measurement error to ensure accurate model evaluation.

TABLE 13. Assessment results of different prediction models.

Models	Performance Metrics				
	Accuracy	Precision	Sensitivity	F1Score	AUC
AdaBoost	0.925	0.935	0.941	0.925	0.925
RF	0.935	0.945	0.937	0.935	0.930
SVM	0.934	0.943	0.936	0.935	0.930
CNN	0.943	0.950	0.943	0.943	0.937
ANN	0.932	0.940	0.933	0.932	0.932
DNN	0.950	0.958	0.951	0.951	0.949
TL-DNN	0.960	0.965	0.955	0.960	0.960

5) EVALUATION METRICS

Accuracy measures the proportion of correct predictions (both true positives and true negatives) across the entire dataset. However, in the context of class imbalance, accuracy alone may not be the best performance metric. Precision evaluates the accuracy of positive predictions, while sensitivity (recall) measures the proportion of actual positives correctly identified, which is crucial in medical applications to minimize missed positive cases. The F1 score, which harmonizes precision and sensitivity, provides a balanced measure of model performance. The Area Under the Curve (AUC) metric evaluates the classifier's performance across various thresholds, with higher AUC values indicating better predictive power.

C. RESULTS

This study attempts to use additional data sets to test whether the methods proposed will differ in different data sets. First, three machine learning methods are used, and then deep learning models are also applied to predict whether hemodialysis patients will be hospitalized again within 30 days, and finally conduct a classification experiment through the transfer learning method is to observe whether there is further improvement in the evaluation metrics.

Finally, a detailed assessment results of different ML and DL models includes AdaBoost, RF, SVM, CNN, ANN, and DNN are shown for comparison in TABLE 13. The performance metrics of AdaBoost and RF models are slightly higher than the results of SVM, it usually depends on the characteristics of the data set. RF generally performs better in stability and handling outliers, while AdaBoost can sometimes provide superior performance when dealing with non-linear data. CNN and DNN models can capture complex patterns in the data, the predictive performance is slightly better than SVM. The performance of ANN depends on its structure and training method, and is slightly lower than CNN and DNN, but can still provide competitiveness in some cases. TL-DNN often achieves excellent results, with the highest predictive performance in this additional dataset.

VI. CONCLUSION

With the coming of the era of big data, AI is gradually becoming mature and put into the medical field. AI-combined medical treatment will be more widely used in the clinical diagnosis and treatment of various diseases, which can shorten the diagnosis time, improve accuracy and reliability, and present data and statistics more quickly to directly and effectively help the clinical treatment and care, avoid overwork and reduce medical costs, and modern medical data collections may have even more category imbalances. In the category of imbalance problems, classifiers will produce errors in training. As a result, it has a very low prediction accuracy for minority class examples. This problem is caused by unbalanced data. In this type of data, the sample number of one class will far exceed that of other categories. When extracting knowledge from unbalanced data, traditional classifiers will pursue high classification accuracy for most classes of samples but have poor prediction accuracy for a few categories. In the medical field, clinical data are often cate-gory imbalanced in classes, and the patients are often a small group of people. Therefore, in order to properly diagnose patients or arrange proper medical treatment, it is necessary to improve the ability to distinguish between a few classes. In this study, data preprocessing and random under-sampling methods were used to compare the evaluation metrics of various types of category imbalance in the medical stroke dataset with the deep learning model. According to the experimental results, the results of evaluation metrics G-Mean, ACC, FNR, and FPR in this study are all better than those in previous studies. This study found that the use of the transfer learning process has improved G-Mean, ACC, and FNR metrics in stroke prediction in comparison with traditional ML models and DL models. This suggests that transfer learning can help diagnose clinical category imbalance data more effectively in the medical field. Verifying that data science models and algorithms work effectively in real-world clinical settings through clinical validation is essential. For researchers, conducting additional clinical validation studies can help con-firm model effectiveness and generalizability. This can be done by collaborating with healthcare professionals, testing on independent datasets, and comparing model performance to existing clinical practices or guidelines.

In this study, under-sampling is adopted to deal with category imbalance. Outlier removal is also inherited to test the improvement of stroke prediction accuracy without validity. In the future, maybe we can try to combine other methods to deal with category unbalanced data sets. In addition, other advanced deep learning models can also be combined to improve the effect of evaluation indicators.

REFERENCES

[1] D. Liang, Q. Guan, M. Huang, Y. He, Y. Ou, M. Chen, X. Zheng, and X. Lin, "Changing trends of disease burden of stroke from 1990 to 2019 and its predictions among the Chinese population," *Frontiers Neurol.*, vol. 14, Oct. 2023, doi: 10.3389/fneur.2023.1255524.

[2] H. Ge, T. Zhou, C. Zhang, Y. Cun, W. Chen, Y. Yang, Q. Zhang, H. Li, J. Zhong, X. Zhang, H. Feng, and R. Hu, "Targeting ASIC1a promotes neural progenitor cell migration and neurogenesis in ischemic stroke," *Research*, vol. 6, Jan. 2023, doi: 10.34133/research.0105.

[3] W. Wang, F. Qi, D. Paul Wipf, C. Cai, T. Yu, Y. Li, Y. Zhang, Z. Yu, and W. Wu, "Sparse Bayesian learning for end-to-end EEG decoding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15632–15649, Dec. 2023, doi: 10.1109/TPAMI.2023.3299568.

[4] M. M. Hassan, M. Raihan, M. H. B. Khan, T. Dhali, M. M. Rahman, and Z. H. Sneha, "A hybrid machine learning approach to predict the risk of having stroke," in *Proc. 2nd Int. Conf. Comput. Advancements*, Mar. 2022.

[5] K.-Y. Chuang, S.-C. Wu, A.-H.-S. Ma, Y.-H. Chen, and C.-L. Wu, "Identifying factors associated with hospital readmissions among stroke patients in Taipei," *J. Nursing Res.*, vol. 13, no. 2, pp. 117–128, Jun. 2005.

[6] A. H. Lee, K. K. W. Yau, and K. Wang, "Recurrent ischaemic stroke hospitalisations: A retrospective cohort study using western Australia linked patient records," *Eur. J. Epidemiol.*, vol. 19, no. 11, pp. 999–1003, Nov. 2004.

[7] S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, S. Cheng, and F. N. Delling, "Heart disease and stroke statistics-2021 update: A report from the American Heart Association," *Circulation*, vol. 143, no. 8, pp. e254–e743, 2021.

[8] A. Di Carlo, *Human and Economic Burden of Stroke*, vol. 38. London, U.K.: Oxford Univ. Press, 2009, pp. 4–5.

[9] Y. Almeida, M. S. Sirsat, S. B. I. Badia, and E. Fermé, "AI-rehab: A framework for AI driven neurorehabilitation training—The profiling challenge," in *Healthinf*, 2020, pp. 845–853.

[10] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, Nov. 2019, Art. no. 101723.

[11] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, Jun. 2022.

[12] S. Cheon, J. Kim, and J. Lim, "The use of deep learning to predict stroke patient mortality," *Int. J. Environ. Res. Public Health*, vol. 16, no. 11, p. 1876, May 2019.

[13] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[14] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[15] F. Kamalov, S. Moussa, and J. A. Reyes, "KDE-based ensemble learning for imbalanced data," *Electronics*, vol. 11, no. 17, p. 2703, Aug. 2022.

[16] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, Jun. 2017.

[17] E. M. Alanazi, A. Abdou, and J. Luo, "Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models," *JMIR Formative Res.*, vol. 5, no. 12, Dec. 2021, Art. no. e23440.

[18] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100116.

[19] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, 2021.

[20] S. Akter, D. Das, R. U. Haque, M. I. Q. Tonmoy, M. R. Hasan, S. Mahjabeen, and M. Ahmed, "AD-CovNet: An exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105657.

[21] C. C. Tusell-Rey, O. Camacho-Nieto, C. Yáñez-Márquez, and Y. Villuendas-Rey, "Customized instance random undersampling to increase knowledge management for multiclass imbalanced data classification," *Sustainability*, vol. 14, no. 21, p. 14398, Nov. 2022, doi: 10.3390/su142114398.

[22] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for improved sentiment analysis via random minority oversampling for user tweet review classification," *Electronics*, vol. 11, no. 19, p. 3058, Sep. 2022.

[23] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[24] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5718–5727, Apr. 2009.

[25] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.

[26] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the category imbalance problem," in *Proc. 4th Int. Conf. Natural Comput.*, 2008.

[27] P. Hart, "The condensed nearest neighbor rule (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 515–516, May 1968.

[28] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Proc. Conf. Artif. Intell. Med. Eur.*, 2001.

[29] I. Tomek, "Two modifications of CNN," Tech. Rep., 1976.

[30] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972.

[31] R. F. A. B. de Morais and G. C. Vasconcelos, "Boosting the performance of over-sampling algorithms through under-sampling the minority class," *Neurocomputing*, vol. 343, pp. 3–18, May 2019.

[32] A. Zughrat, M. Mahfouf, Y. Y. Yang, and S. Thornton, "Support vector machines for class imbalance rail data classification with bootstrapping-based over-sampling and under-sampling," *IFAC Proc. Volumes*, vol. 47, no. 3, pp. 8756–8761, 2014.

[33] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling," *Soft Comput.*, vol. 15, no. 10, pp. 1909–1936, Oct. 2011.

[34] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, Jun. 2015.

[35] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Comput. Appl.*, vol. 28, no. 12, pp. 3941–3951, Dec. 2017.

[36] F. Behrad and M. S. Abadeh, "An overview of deep learning methods for multimodal medical data mining," *Expert Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 117006.

[37] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, and M. Lee, "The skin cancer classification using deep convolutional neural network," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 9909–9924, Apr. 2018.

[38] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[39] H.-T. Nguyen, E.-H. Lee, and S. Lee, "Study on the classification performance of underwater sonar image classification based on convolutional neural networks for detecting a submerged human body," *Sensors*, vol. 20, no. 1, p. 94, Dec. 2019.

[40] A. Ajit, K. Acharya, and A. Samanta, "A review of convolutional neural networks," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (ic-ETITE)*, Feb. 2020, pp. 1–5.

[41] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[42] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharmaceutical Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, Jun. 2000.

[43] H. Kukreja, N. Bharath, C. S. Siddesh, and S. Kuldeep, "An introduction to artificial neural network," *Int. J. Adv. Res. Innov. Ideas. Educ.*, vol. 1, no. 5, pp. 27–30, 2016.

[44] A. D. Dongare, R. R. Kharde, and A. D. Kachare, "Introduction to artificial neural network," *Int. J. Eng. Innov. Technol.*, vol. 2, no. 1, pp. 189–194, 2012.

[45] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[46] A. Lozano-Diez, R. Zazo, D. T. Toledano, and J. Gonzalez-Rodriguez, "An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0182580.

[47] T. Liu, W. Fan, and C. Wu, "Data for A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets," *Mendeley Data, V1*, 2019, doi: 10.17632/x8ygrw87jw.1.

[48] M. Aamir and S. M. A. Zaidi, "DDoS attack detection with feature engineering and machine learning: The framework and performance evaluation," *Int. J. Inf. Secur.*, vol. 18, no. 6, pp. 761–785, Dec. 2019.

[49] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *J. Clin. Epidemiol.*, vol. 151, pp. 185–188, Nov. 2022.

[50] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider, "Missing data in surveys: Key concepts, approaches, and applications," *Res. Social Administ. Pharmacy*, vol. 18, no. 2, pp. 2308–2316, Feb. 2022.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 39, no. 1, pp. 1–22, Sep. 1977.

[52] B. Agbo, H. Al-Aqrabi, R. Hill, and T. Alsboui, "Missing data imputation in the Internet of Things sensor networks," *Future Internet*, vol. 14, no. 5, p. 143, May 2022.

[53] B. Agbo, Y. Qin, and R. Hill, "Best fit missing value imputation (BFMVI) algorithm for incomplete data in the Internet of Things," in *Proc. 5th Int. Conf. Internet Things, Big Data Secur.*, 2020.

[54] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, Apr. 2022.

[55] H. M. Qasim, O. Ata, M. A. Ansari, M. N. Alomary, S. Alghamdi, and M. Almehmadi, "Hybrid feature selection framework for the Parkinson imbalanced dataset prediction problem," *Medicina*, vol. 57, no. 11, p. 1217, Nov. 2021.

[56] V. H. A. Ribeiro and G. Reynoso-Meza, "Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets," *Expert Syst. Appl.*, vol. 147, Jun. 2020, Art. no. 113232.

[57] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," *Cognit. Syst. Res.*, vol. 61, pp. 32–44, Jun. 2020.

[58] W. Koehrsen, "Overfitting vs. underfitting: A complete example," *Towards Data Sci.*, vol. 405, pp. 1–12, Jan. 2018.

[59] Y. Liu, B. Ma, and Y. Wang, "Study on prediction model of stroke risk based on decision tree and regression model," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 4798–4801.

[60] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, Nov. 2018.

[61] J. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," in *Proc. SAS Global Forum*, vol. 12, Apr. 2017, pp. 1–4.

[62] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: Progress in machine intelligence for rational drug discovery," *Drug Discovery Today*, vol. 22, no. 11, pp. 1680–1685, Nov. 2017.

[63] S. A. Alasadi and W. S.Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[64] G. C. O'Connell, A. B. Petrone, M. B. Treadway, C. S. Tennant, N. Lucke-Wold, P. D. Chantler, and T. L. Barr, "Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke," *npj Genomic Med.*, vol. 1, no. 1, pp. 1–9, Nov. 2016.

[65] K. Theofilatos, A. Korfiati, S. Mavroudi, M. C. Cowperthwaite, and M. Shpak, "Discovery of stroke-related blood biomarkers from gene expression network models," *BMC Med. Genomics*, vol. 12, no. 1, pp. 1–15, Dec. 2019.

[66] P. Ren, J.-Y. Wang, H.-L. Chen, X.-W. Lin, Y.-Q. Zhao, W.-Z. Guo, Z.-R. Zeng, and Y.-F. Li, "Diagnostic model constructed by nine inflammation-related genes for diagnosing ischemic stroke and reflecting the condition of immune-related cells," *Frontiers Immunol.*, vol. 13, Dec. 2022.

[67] B. Stamova, G. C. Jickling, B. P. Ander, X. Zhan, D. Liu, R. Turner, C. Ho, J. C. Khoury, C. Bushnell, A. Pancioli, E. C. Jauch, J. P. Broderick, and F. R. Sharp, "Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e102550.

[68] T. L. Barr, Y. Conley, J. Ding, A. Dillman, S. Warach, A. Singleton, and M. Matarin, "Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling," *Neurology*, vol. 75, no. 11, pp. 1009–1014, Sep. 2010.

[69] T. Krug, J. P. Gabriel, R. Taipa, B. V. Fonseca, S. Domingues-Montanari, I. Fernandez-Cadenas, H. Manso, L. O. Gouveia, J. Sobral, I. Albergaria, G. Gaspar, J. Jiménez-Conde, R. Rabionet, J. M. Ferro, J. Montaner, A. M. Vicente, M. R. Silva, I. Matos, G. Lopes, and S. A. Oliveira, "TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches," *J. Cerebral Blood Flow Metabolism*, vol. 32, no. 6, pp. 1061–1072, Jun. 2012.
[70] S. Sheikholeslami, M. Meister, T. Wang, A. H. Payberah, V. Vlassov, and J. Dowling, "AutoAblation: Automated parallel ablation studies for deep learning," in *Proc. 1st Workshop Mach. Learn. Syst.*, Apr. 2021, pp. 55–61.
[71] B. M. Kissela, J. C. Khoury, K. Alwell, C. J. Moomaw, D. Woo, O. Adeoye, M. L. Flaherty, P. Khatri, S. Ferioli, F. De Los Rios La Rosa, J. P. Broderick, and D. O. Kleindorfer, "Age at stroke: Temporal trends in stroke incidence in a large, biracial population," *Neurology*, vol. 79, no. 17, pp. 1781–1787, Oct. 2012.
[72] C.-H. Tsai, D.-H. Shih, J.-H. Tu, T.-W. Wu, M.-G. Tsai, and M.-H. Shih, "Analyzing monthly blood test data to forecast 30-day hospital readmissions among maintenance hemodialysis patients," *J. Clin. Med.*, vol. 13, no. 8, p. 2283, Apr. 2024, doi: 10.3390/jcm13082283.

DONG-HER SHIH received the Ph.D. degree in electrical engineering from National Cheng Kung University, Taiwan, in 1986. He is currently a Senior Professor with the Department of Information Management, National Yunlin University of Science and Technology, Douliu, Yunlin, Taiwan. He has published more than 90 journal articles. His current research interests include data mining, information security, blockchain, healthcare, business intelligence, deep learning, and big data. He is an Associate Editor of *International Journal of Mobile Communication*.

YI-HUEI WU received the M.S. degree from the Department of Information Management, National Yunlin University of Science and Technology, Taiwan, in 2022, where she is currently pursuing the Ph.D. degree. Her current research interests include healthcare and data mining.

TING-WEI WU received the M.S. and Ph.D. degrees from the Department of Information Management, National Yunlin University of Science and Technology, Taiwan, in 2015 and 2021, respectively. He is currently a Postdoctoral Researcher with the National Yunlin University of Science and Technology. His research interests include big data analysis, data mining, deep learning, blockchain, and computer security.

HUEI-YING CHU received the M.S. degree from the Department of Information Management, National Yunlin University of Science and Technology, Douliu, Yunlin, Taiwan, in 2023. Her major research interests include deep learning, healthcare, and data mining.

MING-HUNG SHIH received the M.S. degree in electrical and computer engineering from North Carolina State University, in 2011, and the Ph.D. degree in electrical and computer engineering from Iowa State University, USA, in 2020. His research interests include big data analytics, computer security, digital health, and data mining.