Section A-Research paper
ISSN 2063-5346

Stroke Risk Prediction With Hybrid Deep Transfer Learning Framework

R.Sreedhar
Associate Professor,Departmement of IT
Sridevi Women's Engineering
College,Telangana
rachasreedharswec@gmail.com

Sandugula Shriya
BTech Student,,Departmement of IT
Sridevi Womens Engineering
College,Telangana
shriya61s@gmail.com

Bellary Chaitanya Lakshmi
BTech Student, Departmement of IT
Sridevi Womens Engineering
College,Telangana
bellarychaitanyalakshmi@gmail.com

Dorishetti Srinidhi Varma
BTech Student, Departmement of IT
Sridevi Womens Engineering
College,Telangana
srinidhi.abhi267@gmail.com

ABSTRACT: Stroke has become the world's biggest cause of mortality and long-term disability, with no effective therapy. Deep learning-based techniques may beat current stroke risk prediction algorithms, but they need enormous amounts of well-labeled data. Stroke data is often transferred in tiny bits around multiple institutions due to the strong privacy protection policy in health-care systems. Furthermore, the positive and negative cases of such data are very skewed. Transfer learning may handle minor data issues by using expertise from a related topic, particularly when numerous data sources are available. We present a unique Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) approach in this paper to harness the information structure from many correlated sources (i.e., external stroke data, chronic diseases data, such as hypertension and diabetes). The proposed system has been thoroughly evaluated in both synthetic and real-world contexts, and it beats the best stroke risk prediction algorithms currently available. It also demonstrates the feasibility of real-world deployment across many hospitals using 5 G/B5G infrastructure.

Keywords *– stroke risk prediction, transfer learning, generative adversarial networks, active learning, Bayesian optimization.*

1. INTRODUCTION

S TROKE is one of the most common illnesses that may cause death or long-term impairment in the elderly all around the globe. According to a recent estimate [1,] around 795 000 persons in the United States have a new or recurrent stroke each year; one stroke occurrence happens every 40 seconds. One in every five patients who had a stroke died within a year [2]. The expense of treatment and rehabilitation for the survivors imposes an exceedingly large burden on their families and the health-care system. The direct and indirect cost of stroke occurrences was about 45.5 billion US dollars from 2014 to 2015 [3]. Thus, precise stroke prediction is extremely important so that the expense of early therapies

to postpone the start of and lower the risks of stroke may be reduced. Several papers have been published that use medical data (e.g., electronic health records and retinal images) to create Stroke Risk Prediction (SRP) Models. Classical machine learning techniques [4], [5] (e.g., Support Vector Machine (SVM), Decision Tree, Logistic Regression) and deep learning-based approaches [6]-[11] may be generally classified. Deep neural networks (DNN) have been shown to perform best in stroke prediction [8]. However, one well-known disadvantage is that such a model is dependent on the availability of vast amounts of well-labeled data. In the actual world, the amount of trustworthy data necessary may not be easily accessible [12]. Sharing stroke data across hospitals is often challenging due to the health-care system's rigorous privacy protection policy. As a result, the whole collection of stroke data is often disseminated in discrete portions among many institutions. Furthermore, the positive and negative incidences in stroke data are severely skewed. As a result, DNN-based SRP models may perform badly in real-world deployment [13].

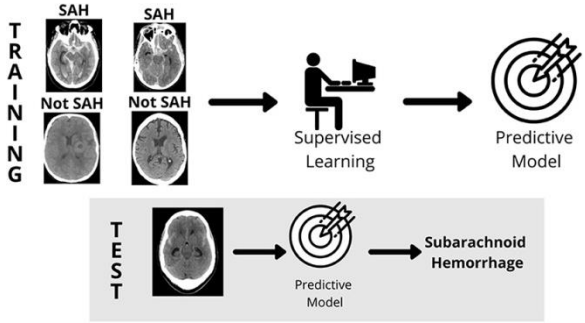

Fig.1: Example figure

Though the stroke data is limited, certain prevalent chronic conditions (e.g., hypertension and diabetes) have sufficiently substantial data and have been shown in clinical trials to be significantly linked with stroke development [14], [15]. Transfer Learning (TL) techniques provide an appropriate framework to solve small data issues when numerous linked sources are available. The majority of extant TL efforts use single transfer methodologies such as feature transfer, instance transfer, and network transfer. A recent study suggested a hybrid adapted-embedding technique and shown experimentally that hybrid transfer beats single transfer methods. In the Meta-learning framework, transfer learning is also employed for low-resource predictive modelling using patient EHRs.

2. LITERATURE REVIEW

Heart disease and stroke statistics—2017 update a report from the American Heart Association:

Each year, the American Heart Association (AHA), in collaboration with the Centers for Disease Control and Prevention, the National Institutes of Health, and other government agencies, compiles the most recent statistics on heart disease, stroke, and the AHA's Life's Simple 7 (Figure1), which include core health behaviours (smoking, physical activity [PA], diet, and weight) and health factors (cholesterol, blood pressure). The Statistical Update is an essential resource for the general public, policymakers, media professionals, doctors, healthcare administrators, researchers, health activists, and anyone looking for the most up-to-date information on these causes and conditions. Cardiovascular disease (CVD) and stroke impose enormous health and economic costs in the United States and across the world. The Update also includes the most recent data on a variety of major clinical heart and circulatory disease conditions and outcomes (including stroke, congenital heart disease, rhythm disorders, subclinical atherosclerosis, coronary heart disease, heart failure (HF), valvular disease, venous disease, and peripheral arterial disease) (including quality of care, procedures, and economic costs). Since 2006, the yearly editions of the Statistical Update have been

mentioned in the literature over 20,000 times. The different Statistical Updates were mentioned over 4000 times in 2015.

An integrated machine learning approach to stroke prediction

Stroke is the third largest cause of mortality in the United States and the leading cause of significant long-term disability. Accurate stroke prediction is critical for early intervention and therapy. On the Cardiovascular Health Study (CHS) dataset, we compare the Cox proportional hazards model with a machine learning technique for stroke prediction. We focus on the common issues of data imputation, feature selection, and prediction in medical datasets. We offer a unique automated feature selection technique based on our suggested heuristic: conservative mean. When combined with Support Vector Machines (SVMs), our proposed feature selection technique outperforms the Cox proportional hazards model and the L1 regularised Cox feature selection algorithm in terms of area under the ROC curve (AUC). We also describe a margin-based censored regression method, which combines the notion of margin-based classifiers with censored regression to attain a higher concordance index than the Cox model. Overall, our method exceeds the existing state-of-the-art in both AUC and concordance index measurements. Furthermore, our analysis has found possible risk variables that standard methodologies have not detected. Our technique may be used to predict clinical outcomes in other illnesses where missing data is frequent and risk factors are poorly known.

Using machine learning to improve the prediction of functional outcome in ischemic stroke patients

Ischemic stroke is the main cause of disability and mortality in people globally. Individual prognosis following a stroke is highly reliant on treatment options made by doctors during the acute period. Several scores, including the ASTRAL, DRAGON, and THRIVE, have been offered as measures to assist clinicians estimate a patient's functional prognosis following a stroke in the previous five years. These are rule-based classifiers that employ characteristics accessible when the patient is admitted to the ER. We use machine learning approaches to the challenge of predicting functional outcome of ischemic stroke patients three months after admission in this research. When employing the characteristics available at admission, we demonstrate that a pure machine learning strategy yields only a modestly better Area Under the ROC Curve (AUC) ($0.808pm 0.085$) than the top score ($0.771pm 0.056$). However, we discovered that by gradually adding characteristics that are accessible at later periods in time, we can considerably boost the AUC to a number greater than 0.90. We conclude that the findings confirm the use of admission scores, but also highlight the need of employing additional characteristics, which necessitate more complex approaches, where feasible.

EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques

Ischemic stroke is a leading cause of mortality and disability in adults across the globe. Ischemic stroke phenotyping is critical for medical research and clinical prognosis because of its widely variable morphologies. However, when the study population is huge, this job is not easy. In prior investigations, phenotyping of ischemic stroke was mostly based on human annotation of medical data. Based on structured and unstructured data from electronic medical records, this study investigated alternative methodologies for automated phenotyping of ischemic stroke into the four subtypes of the

Oxfordshire Community Stroke Project categorization (EMRs). A total of 4640 adult patients admitted to a teaching hospital for acute ischemic stroke were included in the study. Aside from the structured elements in the National Institutes of Health Stroke Scale, unstructured clinical narratives were preprocessed using MetaMap to identify medical ideas, which were subsequently encoded into feature vectors. Classifiers were created using a variety of supervised machine learning methods. The study's findings suggest that when paired with structured data, textual information from EMRs might aid in the phenotyping of ischemic stroke. Furthermore, decomposing this multi-class issue into binary classification tasks followed by classification result aggregation might increase performance.

Feature isolation for hypothesis testing in retinal imaging: An ischemic stroke prediction case study

Ischemic stroke is a prominent cause of mortality and disability that is difficult to anticipate. Because of its non-invasiveness and the resemblance between retinal and cerebral microcirculations, retinal fundus photography has been advocated for stroke risk assessment, with previous research suggesting a link between venular calibre and stroke risk. However, other retinal properties may be more suited. Extensive deep learning studies on six retinal datasets are discussed in this publication. The use of segmented vascular tree pictures for feature isolation is used to determine the efficacy of vessel diameter and form alone for stroke classification, and dataset ablation is used to explore model generalizability on unknown sources. The findings imply that vascular diameter and shape may be indicators of ischemic stroke, and that source-specific factors may impact model performance.

3. METHODOLOGY

Deep neural networks (DNN) are said to have the greatest performance in stroke prediction. However, one well-known disadvantage is that such a model is dependent on the availability of vast amounts of well-labeled data. In the actual world, the amount of trustworthy data necessary may not be easily accessible. Sharing stroke data across hospitals is often challenging due to the health-care system's rigorous privacy protection policy. As a result, the whole collection of stroke data is often disseminated in discrete portions among many institutions. Furthermore, the positive and negative incidences in stroke data are severely skewed. As a result, DNN-based SRP models may perform badly in real-world deployment.

Disadvantages:

1. Sharing stroke data across hospitals is often challenging due to the health-care system's rigorous privacy protection policy.
2. DNN-based SRP models may not perform well in real-world deployment.

We present a unique Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) approach in this paper to harness the information structure from many correlated sources (i.e., external stroke data, chronic diseases data, such as hypertension and diabetes). The proposed system has been thoroughly tested in both synthetic and real-world contexts, and it outperforms current stroke risk prediction methods. It also demonstrates the feasibility of real-world deployment across many hospitals using 5G/B5G infrastructure.

Advantages:

1. The suggested framework has a higher capacity to create SRP models.
2. Bayesian Optimization (BO) is a method for model-based global optimization of blackbox functions, with

the Gaussian process being the most often used model owing to its simplicity and flexibility in generating a probabilistic model of the objective function.

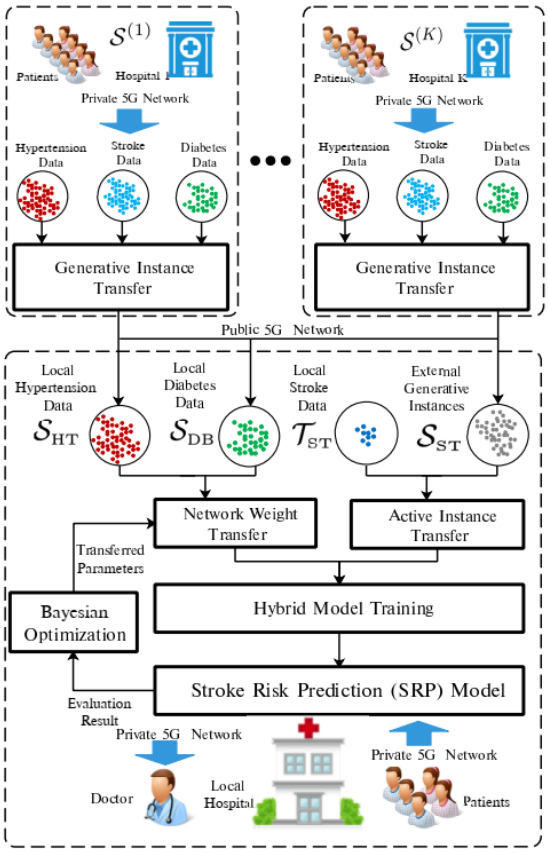

Fig.2: System architecture

MODULES:

To carry out the aforementioned project, we created the modules listed below.

- Data exploration: we will put data into the system using this module.
- Processing: we will read data for processing using this module.
- Using this module, data will be split into train and test.
- Model generation: Building the model - Hybrid Deep TL - DNN - CNN + LSTM - SVM - DT - RF - Voting Classifier - (Naive Bayes + RF + DT). Calculated algorithm accuracy.
- User registration and login: Using this module will result in registration and login.
- Using this module will provide input for prediction.
- Prediction: final predicted shown

4. IMPLEMENTATION

ALGORITHMS:

Hybrid Deep TL: A Hybrid Deep Transfer Learning (HDTL) technique that transfers knowledge structure from various source domains scattered across different hospitals to the target domain of stroke. The proposed HDTL-SRP infrastructure operates in a dispersed manner, requiring no direct sharing of patient information across institutions. It is made up of three parts: (1) Generative Instance Transfer (GIT), which uses GAN in external data to generate synthetic instances for model training, (2) Network Weight Transfer (NWT), which uses data from highly correlated diseases (i.e., hypertension or diabetes), (3) Bayesian Optimization (BO), which finds the best transferred parameters, and (4) Active Instance Transfer (AIT), which selects more informative synthetic stroke instances to create a balanced stroke dataset, which is then used to train models.

DNN: Deep Learning is concerned with the training of massive neural networks with sophisticated input-output transformations. One current use of DL is the mapping of a photo to the name of the person(s) in the shot, like they do on social networks. Another new application of DL is characterising an image with a word.

CNN + LSTM: A CNN-LSTM model is made up of CNN layers that extract features from input data and LSTM layers that forecast sequences. In general, the CNN-LSTM is utilised for activity identification, picture labelling, and video labelling.

SVM: Support Vector Machine (SVM) is a supervised machine learning technique that may be used for classification and regression. Though we call them regression issues, they are best suited for categorization. The SVM algorithm's goal is to identify a hyperplane in an N-dimensional space that clearly classifies the input points.

DT: A decision tree (DT) is a non-parametric supervised learning technique that may be used for classification and regression applications. It has a tree structure that is hierarchical and consists of a root node, branches, internal nodes, and leaf nodes.

RF: A Random Forest Method (RF) is a supervised machine learning algorithm that is widely used in Machine Learning for classification and regression issues. We know that a forest is made up of many trees, and the more trees there are, the more vigorous the forest is.

Voting Classifier: Voting Classifier is a machine-learning method that Kagglers often employ to improve the performance of their model and move up the rank ladder. Voting Classifier may also be used to increase performance on real-world datasets, although it has significant limitations.

5. EXPERIMENTAL RESULTS

Fig.3: Home screen

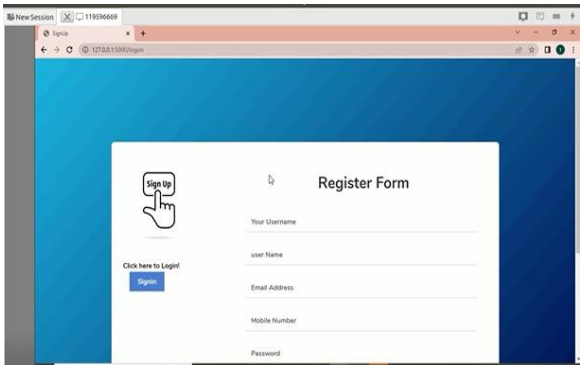

Fig.4: User registration

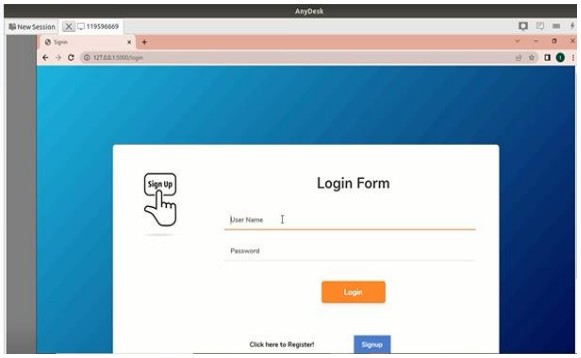

Fig.5: user login

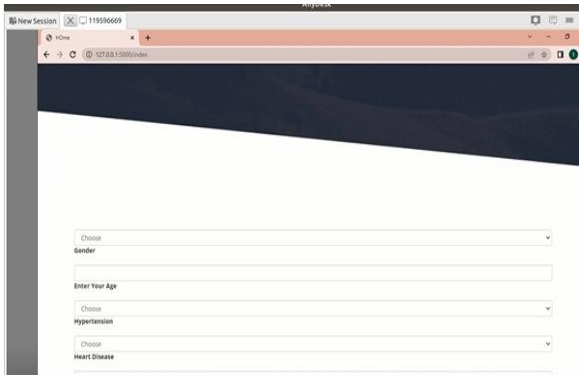

Fig.6: Main screen

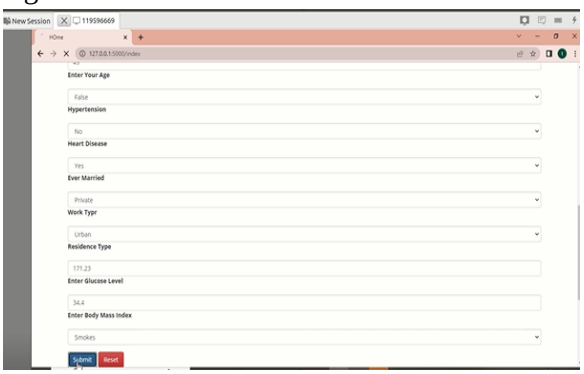

Fig.7: User input

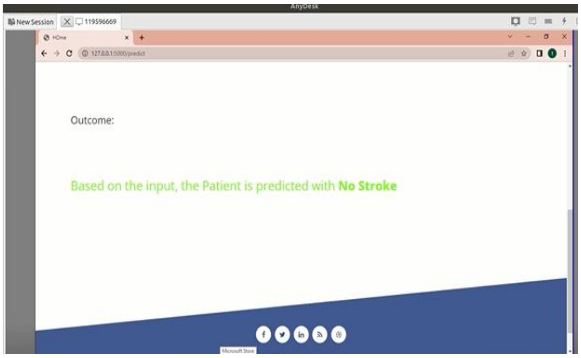

Fig.8: Prediction result

6. CONCLUSION

This effort solved the problems associated with SRP with little and unbalanced stroke data. We proposed a novel Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) framework comprised of three key components: (1) Generative Instance Transfer (GIT) for utilising external stroke data distribution among multiple hospitals while maintaining privacy, (2) Network Weight Transfer (NWT) for utilising data from highly correlated diseases (i.e., hypertension or diabetes), and (3) Active Instance Transfer (AIT) for balancing the stroking. In both synthetic and real-world contexts, the proposed HDTL-SRP framework outperforms state-of-the-art SRP models. There are several open questions for future research, including how to (1) extend NWT to consider multiple chronic diseases simultaneously, (2) learn the optimal number of layers to be transferred automatically, (3) implement the system of other diseases as the health-care data share similar characteristics (i.e., small and imbalanced), and (4) improve the interpretability of the SRP model as the interpretable mechanism.

REFERENCES

[1] E. J. Benjamin, M. J. Blaha, and S. E. Chiuve, "Heart disease and stroke statistics—2017 update a report from the American Heart Association," Circulation, vol. 135, no. 10, pp. e146—e603, 2017.

[2] S. Koton et al., "Stroke incidence and mortality trends in US communities, 1987 to 2011," JAMA, vol. 312, no. 3, pp. 259–268, 2014.

[3] E. J. Benjamin, P.Muntner, andM. S. Bittencourt, "Heart disease and stroke statistics—2019 update: A report from the American Heart Association," Circulation, vol. 139, no. 10, pp. e56–e528, 2019.

[4] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2010, pp. 183–192.

[5] M. Monteiro et al., "Using machine learning to improve the prediction of functional outcome in ischemic stroke patients," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 15, no. 6, pp. 1953–1959, Nov./Dec. 2018.

[6] S. F. Sung, C. Y. Lin, and Y. H. Hu, "EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques," IEEE J. Biomed. Health Inform., vol. 24, no. 10, pp. 2922–2931, Oct. 2020.

[7] G. Lim et al., "Feature isolation for hypothesis testing in retinal imaging: An ischemic stroke prediction case study," in Proc. AAAI Conf. Artif. Intell., 2019, vol. 33, pp. 9510–9515.

[8] 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, 2019, Art. no. 1876.

[9] D. R. Pereira, P. P. R. Filho, G. H. de Rosa, J. P. Papa, and V. H. C. de Albuquerque, "Stroke lesion detection using convolutional neural networks," in Proc. Int. Joint Conf. Neural Netw., 2018, pp. 1–6.

[10] D. Teoh, "Towards stroke prediction using electronic health records," BMC Med. Informat. Decis. Mak., vol. 18, no. 1, pp. 1–11, 2018.

[11] 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, 2019, Art. no. 101723.
[12] F. Wang, L. P. Casalino, and D. Khullar, "Deep learning in medicine— promise, progress, and challenges," JAMA Intern. Med., vol. 179, no. 3, pp. 293–294, 2019.
[13] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," inProc. IEEE Int. Conf. Comput. Vis., 2017, pp. 843–852.
[14] A. O'Brien, C. Rajkumar, and C. J. Bulpitt, "Blood pressure lowering for the primary and secondary prevention of stroke: Treatment of hypertension reduces the risk of stroke," J. Cardiovasc. Risk, vol. 6, no. 4, pp. 203–205, 1999.
[15] J. E. Manson et al., "A prospective study of maturity-onset diabetes mellitus and risk of coronary heart disease and stroke in women," Arch. Intern. Med., vol. 151, no. 6, pp. 1141–1147, 1991.