imbalance data. Section IV explains results and the last section concludes the summary of the work.

II. RELATED WORKS

The proposed model is based on IoT (Internet of Things) which take various sensors data, distribute it intelligently and execute desired commands of paralyzed patients for their caretakers to help them support and render therapy by using AI-IoT-SES (Artificial Intelligence (AI)-powered Smart and Lightweight Exoskeleton System) [5]. Due to intestinal blockage in the absence of a physical obstruction paralytic ileus occur, author forecast mortality in PI (Paralytic ileus) patients by using SRML (Statistically Robust Machine Learning)-Mortality Predictor framework with Support Vector Machine (using RBF kernel) and achieve accuracy around 81.38% [6]. Using kinematic coefficients calculated from the joint model and machine learning, propose a quantitative evaluation approach for wrist paralysis in stroke patients. The computed stiffness's and viscosities were used to classify the patients into three separate clusters using a Support Vector Machine and a K-means approach.

It is found that the proposed HDTL-SRP framework outperforms the state-of-the-art SRP models in both synthetic and real-world scenarios [7]. The two coefficients not only helped distinguish patients based on their Brunnstrom stage, but they also demonstrated that patients in the same stage could be classified more precisely [8].

A system that intelligently evaluates facial damage in 43 testing videos from 14 participants is shown here. The system's average mean squared error is 1.6 per cent when using an Artificial Neural Network. There are 16 recovery sequences for patients who have improved from onset to recovery, as well as 22 post-operative single session movies. Although the color information was considered throughout the project as a tool to better the feature location and stabilization steps, the movies themselves were shot in color and then converted to black and white [9]. The asymmetry between two sides of the face was measured using P face, which is derived from D face, and the expression variations between the patient and normal participants were measured using eigenflow. The support vector machine then merged the findings from P face and eigenflow to obtain a P degree. " P degree was found to be able to discern between paralysis states ($P \geq 0$) and normal states (P degree < 0) in a research of 25 people, with the ability to grade facial paralysis automatically [10].

The Relief-F feature selection (FS) algorithm is used to choose the best relevant feature for the deep learning classifier using the hybrid Feature Extraction method, which has the advantage of extracting relevant features from the EMG (Electromyography)signals used to measure electrical currents generated in muscles during its contraction representing neuromuscular activities. On the total EMG signal, the suggested hybrid FE technique achieved 88% precision, compared to 65% precision for the existing neural network (NN) and 35% precision for the support vector machine (SVM) [11]. Because of the various abnormalities in spellings, classifying text from article tagging in Roman-Urdu is a challenging operation. For example, the term khubsurat (beautiful) has many spellings present in Roman-Urdu. For compression different Machine Learning algorithms are used in the experiment in which the SGD classifier, our model predicts the best outcome for identifying the desired class, with a 93.50% accuracy rate [12]. Due to impairment for either side of the body [13], the quality of life is affected by stroke survivors most [14].

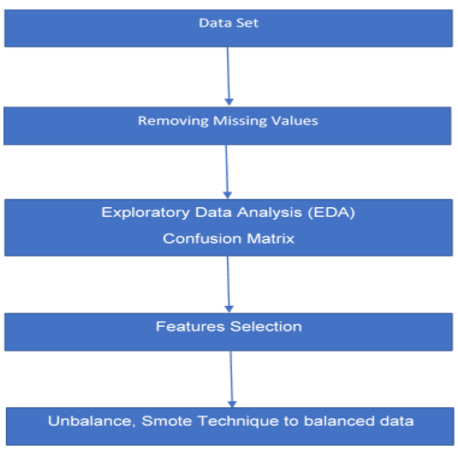
In many cases, the physical disability of stroke survivors is around three months which is closely observed in 35% of stroke survivors suffering paralysis of the leg and 25% of total survivors are not comfortable walking without any assistance [15]. 747, 514 participants of China National Stroke Screen Data of the year 2017 is taken and apply different machine learning algorithms to classify the stroke risk based on recall and precision, 99.94% boosting model with decision tree and 97.33% model based on the random forest were analyzed [16]. Further, the 2006 stroke patient's data from Taiwan Stroke Registry (TSR) have been studied with different machine learning models such as Artificial Neural Network (ANN), Support Vector Machine (SVM), and random forest (RF), including Hybrid Artificial Neural Network (HANN) and found 97% accuracy for 90-days stroke outcome predictions in both ischemic and hemorrhagic stroke with hyper-parameters of 10-time repeated hold-out with 10-fold cross-validation [17].

The German statutory health insurance company data were analyzed for the two time periods 2006 to 2008, and 2014 to 2016, and the results revealed that the occurrence and death risk of stroke patients decreased in both men and women regardless of income groups. Whereas, the men having higher incomes, decreasing incidences and improvements in mortality stroke-free lifetime could be gained but its effected more on lower-income group leading to decrease in life years. On the other hand, the same observation can be applied to women with and found no significant changes in the life years [18]. The income level of different countries has been measured, ranging from 94 to 117 figure 100,000 person-years to capture the incidence of stroke which is very common in low-income countries than high-income countries [19]. The binary classification model applied on face photographs to identify facial paralysis based on facial landmark extraction, facial measure computation, and facial paralysis classification by using a multi-layer perceptron approach to identify asymmetry levels within the face elements output label [20]. To prevent overfitting, which mostly occurs in CNNs (Convolutional Neural Networks), GAN (Generative Adversarial Network) is applied on large dataset face images to extend the training dataset that extracts palsy-specific features. Palsy disease categories into five benchmarked grades by extracting deeply learn features [21]. House-Brackmann (HB) classification techniques are used on 80 facial paralyses patient images for analysis and the facial blood flow of FP patients were measured to extract the facial blood flow distribution characteristics by dividing the faces into concerned regions. The purposed method is achieved a 97.14% accuracy score that outperforms the state-of-the-art systems [22]. Several studies showed the existing performance used to measure and validate the life of quality and reported outcome from facial paralysis patients throughout the population and significant reductions were found in the false-negative rate. Only 19.1% were predicted cerebral stroke for clinical diagnosis data

having 43400 records of potential patients which includes 783 occurrences of stroke by applying a hybrid machine learning approach with incompleteness and class imbalance [23].
The RF algorithm is successfully used to identify the missing values before classification and then to predict stroke on an imbalanced dataset by automated hyperparameter optimization (Auto-HPO), based on DNN [24],[25].

III. METHODOLOGY

The dataset that used in the experiment is freely available on the website [26], for predicting stroke events have both numerical and categorical variables having 5110 entries and 10 clinical input features with 1 target variable named ‘stroke’ with two inputs 0 (no-stroke) and 1 (stroke) for the prediction based on input features i.e., smoking status, age, gender, and various diseases with relevant information. The target variable is coded as 1 for positive cases (has a stroke) and 0 for negative cases (does not have a stroke). Randomly selected data use for training and testing features by applying over-sampling and under-sampling techniques.



Missing Data

From the given dataset firstly identify any duplicate value which is not found, next step is to find the missing value which appears in BMI (body mass index) column. Figure1 show missing values appear.

Exploratory Data Analysis (EDA)

Correlation

Correlation is computed using a correlation matrix on stroke data shown in figure 2. Whereas, the data were normalized according to equation-1.

$$f_n = \left\{ \frac{f_o - \min(F)}{\max(F) - \min(F)} \right\} \qquad (1)$$

where f_n is the updated features, max(F) and min(F) are representing the maximum and minimum values of the features.

To identify the correlation between the input variable and target variable Pearson correlation has been applied, -1 indicate negative, 0 shows no and 1 display total positive correlation between variables. Here in figure 3 greater than 0.1 is highly correlated with target variable i.e. age, hypertension, heart disease and average glucose level.

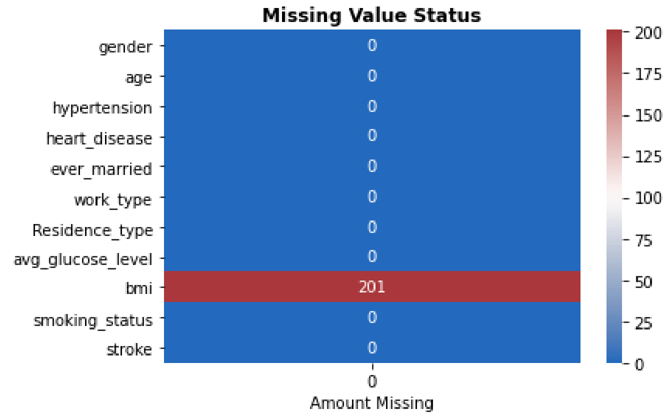


Fig. 1 Chart of Missing Values

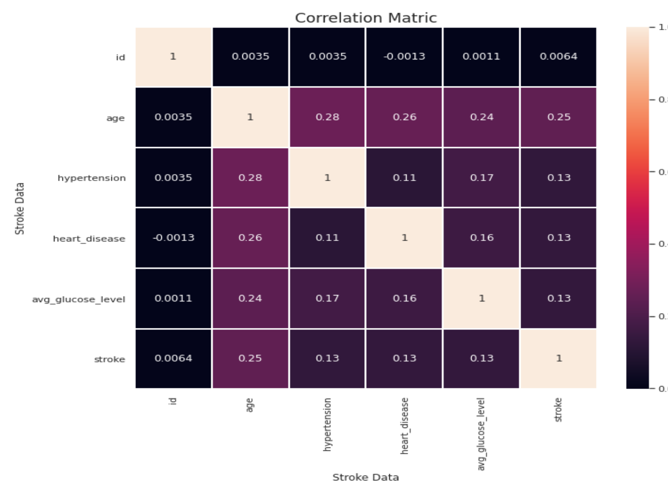


Fig. 2 Chart of Missing Values

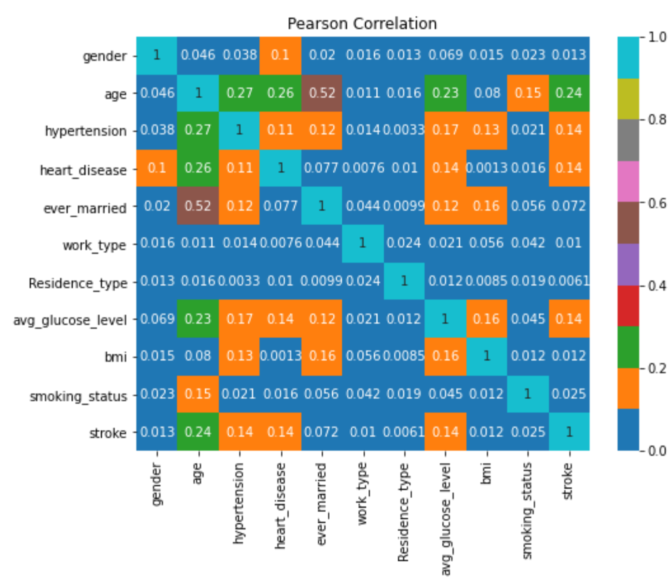


Fig. 3 Pearson Correlation Matrix

Analyzing Patterns using Visualization

There are a few other graphs that compute for analysis and feature selection. Figure 4 shows that patients married status either Yes or No in both conditions, male patients' response is higher than female. The variation in dataset visible on work type of all Patients feature based on their ages, the ratio of female patients is in private job is high like self-employed and Government job, there is very little variation in children and Never Employed type.

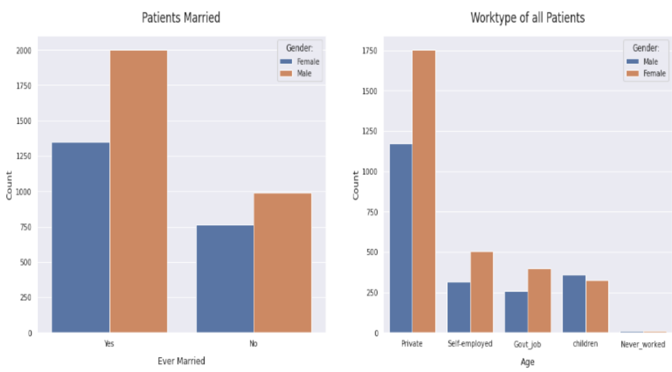


Fig. 4 Patients Married and Work Type

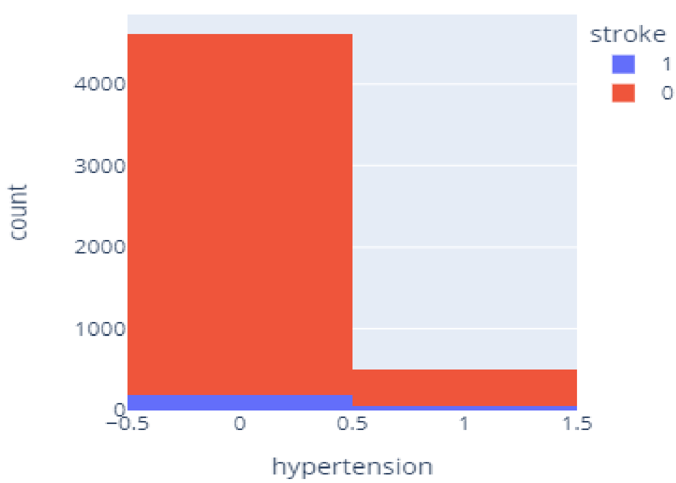


Fig. 5 Mean of Hypertension and Stroke

Stroke probability for those who have hypertension is quite different than for those who don't. 13.2% and 3.9% respectively It means that people with hypertension are almost 3.3 times more likely to get stroke than the ones who don't have hypertension.

Feature Selection.

Dataset used in this experiment is imbalanced data, 95% of the target variable is 'No Stroke' with 4861 records and 249 is found with 'Stroke' class. Before applying machine, learning algorithms t needs to be pre-processed, so SMOTE technique is applied to increase the data in minority class which generate a new sample by using the oversampling method and can avoid over-fitting data. To reduce the difference between both classes under-sampling method is used to generate random sample data in the majority class.

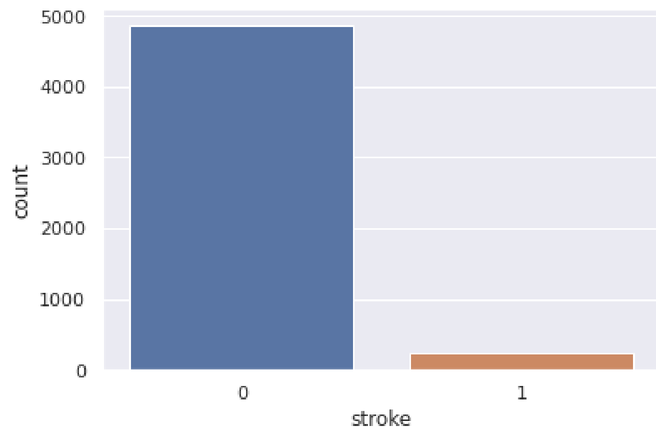


Fig. 6 Unbalanced Data

IV. Results and Discussion

The analysis result is based on Imbalanced data followed by Over-sampling and under-sampling using SMOTE technique for balancing minority and majority class for classification. Precision and recall are used to find the ratio of truly positive with several positive items result after classification and truly positive count in the classification found an entire dataset truly positive items. By taking precision and recall at the same time F1-score is also measured lastly Receiver Operating Characteristic (ROC) curve value is computed for the model's accuracy prediction.

$$P_r = \frac{TP}{TP+FP} \tag{2}$$

$$= \frac{TP}{TPP} \tag{3}$$

$$R_c = \frac{TP}{TP+FN} \tag{4}$$

$$= \frac{TP}{TAP} \tag{5}$$

$$F_1 Score = \frac{P_r R_c}{P_r + R_c} \times 2 \tag{6}$$

Equation (5) is the generalized form of recall for the true positive (TP) and the total actual positive (TAP), whereas, equation (6) is used to balance the precision and recall. The average precision was calculated in equation (7) to summarize the weighted mean precision at the threshold.

$$AP_r = \sum_n \left(P_r - P_{r-1}\right) P_n \tag{7}$$

Case1

The sample size of the target variable is unbalanced having 4861 samples from 0 negative class, and 249 samples are related to the positive class mentioned as labelled 1. 0.15% (15%) as test set and left 85% as a training set, the evolution results of each model using unbalance data were presented in Table 1. The six learning methods (LR, KNN, SVC, DT, RF and XGB) were analyzed for accuracy of training and testing, Recall precession, F1 score and Roc-accuracy score. Logistic regression score was found to be 85.542 followed by XGB as 85.347 (Table 1), the Logistic Regression algorithm is found to be good as compare to other.

Table 1 Evolution results of each model using Unbalanced Data.

Learning Method	Accuracy (Train)	Accuracy (Test)	Recall	Precision	F1 Score	Roc-Acc Score
Logistic Regression	92.715	92.715	92.715	93.142	92.715	97.549
K-NN Classifier	95.187	95.873	95.873	95.873	95.873	97.813
SVC	93.625	92.571	92.571	92.571	92.571	97.520
Decision Tree Classifier	93.713	93.695	93.695	93.801	93.695	93.752
Random Forest Classifier	96.820	96.837	96.837	96.837	96.837	98.611
XGB-Classifier	94.542	95.632	95.632	95.632	95.632	**98.934**

Table 2 Evolution results of each model using Smote (Upsampling)

Learning Method	Accuracy (Train)	Accuracy (Test)	Recall	Precision	F1 Score	Roc-Acc Score
Logistic Regression	92.715	92.715	92.715	93.142	92.715	97.549
K-NN Classifier	95.187	95.873	95.873	95.873	95.873	97.813
SVC	93.625	92.571	92.571	92.571	92.571	97.520
Decision Tree Classifier	93.713	93.695	93.695	93.801	93.695	93.752
Random Forest Classifier	96.820	96.837	96.837	96.837	96.837	98.611
XGB-Classifier	94.542	95.632	95.632	95.632	95.632	**98.934**

Table 3 Evolution results of each model using Clustercentroids (Downsampling)

Learning Method	Accuracy (Train)	Accuracy (Test)	Recall	Precision	F1 Score	Roc Acc Score
Logistic Regression	98.671	98.518	98.518	98.518	98.518	98.762
K-Neighbors Classifier	95.503	97.912	97.912	97.912	97.912	**99.877**
SVC	98.739	98.820	98.820	98.820	98.820	98.665
Decision Tree Classifier	95.549	98.733	98.733	98.733	98.733	98.720
Random Forest Classifier	96.822	97.778	97.778	97.778	97.778	**99.989**
XGB-Classifier	96.909	98.691	98.691	98.691	98.691	**99.971**

Case 2

Evolution results of each model using Smote (Upsampling technique) is used to balance both classes by taking minority class synthetically generated data points introduce into the dataset. Bias is added into the system prevents the model from inclining towards the majority class because of the additional information. By taking 75% as training and 25% as a testing sample which shows the result of almost 99% Random forest classifier and XGB classifier is good for upsampled data (Table 2). In contrast, the similar study was conducted by Hayder and co-worker among the Parkinson patients by using different ML classifiers and reported that the optimal evaluation metrics using KNN model found to be 87% with 95% accuracy and precision. Whereas, the SVM produced 98.22% accuracy [27].

Case 3

Evolution results of each model using Clustercentroids (Downsampling) were presented in table 3. The majority class to the minority class size with a random subsampling. Random undersampling involves randomly selecting examples from the majority class and deleting them from the training dataset. Both classes have balanced data 249 sample sizes of negative and positive data. Here all algorithms work well with the selection of 0.20% (20%) test size and rest for training, still, data is very less to make it interpretable for future unseen data. The Roc-Accuracy Score of RF is found to be high followed by K-NN and XGB classifier respectively as 99.898, 99.971 and 99.877.

V. Conclusion

In this research, the selected dataset has 10 input features and 1 target (output) feature named "Stroke" use for binary classification. Developed models based on effective machine learning algorithms on imbalanced, over-sample, and under-sample data which shows Logistic Regression with 86% accuracy rate is good on imbalanced data, while Random forest classifier and XGB classifier is excellent on the oversampling mechanism with 99% accuracy rate, lastly almost all algorithms give almost same result while using Cluster-centroids (downsampling).

References

[1] Alqarni, A. J. Rambely, A. S., Alharbi, S. A., & Hashim, I. (2021). Dynamic behaviour and stabilization of brain cell reconstitution after stroke under the proliferation and differentiation processes for stem cells. *Mathematical Biosciences and Engineering*, 18(5), 6288 - 6304. https:// doi.org/10.3934/mbe. 2021314.

[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] Orlowski, P., O'Neill, D., Grau, V., Ventikos, Y. & Payne, S. (2013). Modelling of the physiological response of the brain to ischaemic stroke, *Interface Focus*, 3(2), https://doi.org/10.1098/rsfs.2012.0079.

[4] Hasin, F., Hasan, AKM. B., Islam, AFM. R., Hossain, M. M. & Musa, H. (2018). Reversing Paralysis Technology – Remarkable Breakthrough for the Treatment of Incurable Paralysis, *Int. J. Hum. Technol. Interact. Act.*, 2(1), 83–96.

[5] Kellmeyer, P. (2021). Artificial intelligence in basic and clinical neuroscience: Opportunities and ethical challenges, *Neuroforum*, 25(4), 241–250. https://doi.org/10.1515/nf-2019-0018.

[6] Jacob, S., Alagirisamy, M., Xi, C., Balasubramanian, V., Srinivasan, R., Parvathi, R., Jhanjhi, N. Z. & Islam, S. M. N. (2021). AI and IoT-Enabled Smart Exoskeleton System for Rehabilitation of Paralyzed People in Connected Communities, *IEEE Access*, 9, https://doi.org/10.1109/ ACCESS. 2021. 3083093.

[7] Jie, C., Yingru, C., Jianqiang, L., Jia, W., Zijie, L., Ashoke, K. N. (2021) Stroke Risk Prediction with Hybrid Deep Transfer Learning Framework, *IEEE Journal of Biomedical and Health Informatics*.99, 1-1. https://doi.org/10.1109/JBHI.2021.3088750

[8] Ahmad, F. S., Ali, L., Mustafa, R. U., Khattak, H. A., Hameed, T., Wajahat, I., Kadry, S. & Bukhari, S.A.C. (2021). A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs), *J. Ambient Intell. Humaniz. Comput.*, 12(3), 3283–3293. https://doi.org/10.1007/s12652-020-02456-3.

[9] Kim, J., Park, W. & Kim, J. (2020). Quantitative evaluation of stroke patients' wrist paralysis by estimation of kinematic coefficients and machine learning. *Sensors and Materials*. 32(3),981-990.https://doi.org/10.18494/SAM. 2020 .2628.

[10] McGrenary, S., O'Reilly, B. F. & Soraghan, J. J. (2005). Objective grading of facial paralysis using artificial intelligence analysis of video data, *Proc. - IEEE Symp. Comput. Med. Syst.*, 587–592, https://doi.org// 10.1109 /CBMS.2005.78.

[11] Wang, S., Li, H., Qi, F. & Zhao, Y. (2004). Objective facial paralysis grading based on Pface and eigenflow, *Med. Biol. Eng. Comput.* 42(5), 598–603, https://doi. org/ 10.1 007/BF02347540.

[12] Subramani, P., Srinivas, K., Kavitha, R.B., Sujatha, R. & Parameshachari, B.D. (2021). Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients, *Pers. Ubiquitous Comput.*, https://doi.org/10.1007/s00779-021-01531-6.

[13] Hassan, S. M., Ali, F., Wasi, S., Javed, S., Hussain, I. & Ashraf, S.N. (2019). Roman-Urdu News Headline Clas sification with IR Models using Machine Learning Algorithms. *Indian Journal of Science and Technology*, 12(35), 1-9. https://doi.org/10.17485/ijst/2019/v12i35/146571.

[14] Rondina, J. M., Park, C. H. & Ward, N. S. (2017). Brain regions important for recovery after severe post-stroke upper limb paresis, *J. Neurol., Neuro- surgery Psychiatry*, 88(9), 737–743.https://doi.org/10.1136/jnnp-2016-315030.

[15] Lai, S. M., Studenski,S. P., Duncan, W. & Perera, S. (2002). Persisting consequences of stroke measured by the stroke impact scale, *Stroke*, 33(7), 1840–1844. https:/doi.org/10.1161/01.str.0000019289.15440.f2.

[16] Hendricks, H. T., van Limbeek, J., Geurts, A. C. & Zwarts, M. J. (2002). Motor recovery after stroke: A systematic review of the literature, *Arch. Phys. Med. Rehabil*, 83(11), 1629–1637. https://doi.org/ 10.1053/apmr.2002.35473.

[17] Li, X., Bian, D., Yu, J., Mao, H., Li, M. & Zhao, D. (2019). Using machine learning models to classify stroke risk level based on national screening data, *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 2, 1386–1390. https://doi.org/10.1109/EMBC.2019.8857657.

[18] Lin, C. H., Hsu, K.C., Johson, K.R., Fann, Y.C., Tsai, C.H.,
Li, Y.S., Lien, M., Chang, W.L., Chen, P.L., Lin, C.L. & Hsu, Y. (2020). Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry, *Computer Methods and Programs in Biomedicine*.190, 105381. https://doi.org/ 10.1016/j.cmpb.2020.105381.

[19] Tetzlaff, J., Geyer, S., Tetzlaff, F. & Epping, J. (2020). Income inequalities in stroke incidence and mortality: Trends in stroke-free and stroke-affected life years based on German health insurance data," *PLoS One*, 15(1), 1–13. https://doi.org/10.1371/journal.pone.0227541.

[20] Feigin, V. L., Lawes, C. M., Bennett, D. A. S., Barker-Collo, L. & Parag, V. (2009). Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review, *Lancet Neurol*, 8(4), 355–369. https://doi.org/10.1016/S1474-4422(09)70025-0.

[21] Parra-Dominguez, G. S., Sanchez-Yanez, R. E. & Garcia-Capulin, C. H. (2021). Facial paralysis detection on images using key point analysis, *Appl. Sci.*, 11(5), 2435. https://doi.org/10.3390/app11052435

[22] Sajid, M., Shafique, T., Baig, M. J. A., Riaz, I., Amin, S. & Manzoor, S. (2018). Automatic grading of palsy using asymmetrical facial features: A study complemented by new solutions," *Symmetry (Basel)*, 10(7), 242. http s://doi.org/10 .33 90/sym10070242.

[23] Jiang, C., Wu, J., Zhong, W., Wei, M., Tong, J., Yu, H. & Wang, L. (2020). Automatic Facial Paralysis Assessment via Computational Image Analysis, *J. Healthc. Eng*, 1-10. https://doi.org /10.1155 /2020 /2398542.

[24] Ho, A. L., Scott, A. M., Klassen, A. F., Cano, S. J. A., Pusic, L. & Van Laeken, N. (2012). Measuring quality of life and patient satisfaction in facial paralysis patients: A systematic review of patient-reported outcome measures," *Plast. Reconstr. Surg.*, 130(1), 91–99. https://doi.org/ 10.1097 /PRS.0b013e318 254b08d.

[25] Liu, T., Fan, W. & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset, *Artif. Intell. Med.*, 101p. 101723, https://doi.org/10.1016 /j.artmed.2019.101723.

[26]https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[27] Hayder, M., Q, Oguz, A., Mohammad, A.A., Mohammad, N.A., Saad, A. & Mazen, A. (2021). Hybrid Feature Selection Framework for the Parkinson Imbalanced Dataset Prediction Problem, *medicina*, 57,1217p 1-13, https://doi.org/10.3390/medicina57111217

About the authors:

Syed Muhammad Hassan is a PhD scholar in Computer Science and his area of research is Artificial Intelligence and computing. Mr. Hassan is lecturer at Sindh Madressatul Islam University in Artificial Intelligence department, he has almost 10 years of experience in educational and industry research work. He obtained MPHIL degree in Telecommunication from IQRA University Pakistan.

Syed Asif Ali is obtained his Ph.D. in Computer Science from University of Karachi in 2011. He has been involved in teaching and research for more than a decade. He is a senior member of IACSIT, CSTA , ISTE and University Syndicate.

Basit Hassan holds a Master of Information Systems from the University of Melbourne, Australia, and a Master of Science in Computer Science from Iqra University, Karachi. He has been teaching at tertiary level a lecturer since 2012. Currently he is serving at Sindh Madressatul Islam University since 2015. Before that he served at private universities as full time as well as adjunct faculty. His experience also includes a brief two years at Algorithm Consulting (Pvt) Limited as a developer for AlgoERP an Enterprise Resource Planning for Apparel Industry.

Imtiaz Husain received his master's and a Ph.D. degree in Mathematical Sciences from the University of Karachi, Pakistan in 2008 and 2014 respectively. He is currently pursuing his Ph.D. in Artificial Intelligence. His research focuses on the application of AI-based tools and techniques in healthcare. He is an Assistant Professor in the Department of Artificial Intelligence and Mathematical Sciences Sindh Madressatul Islam University Karachi.

Maaz Rafiq is ding MPHIL in Computer Science from Mohammad Ali Jinnah University and have more than 4 years of expereicen in the field of AI, Machine Learning and Computer Vision. He did his BS Computer Science from Mohammad Ali Jinnah University.

Saeed Ahmed obtained his master’s degree in Telecommunication offered in Department of Computer Sciences of IQRA University Karachi, Pakistan in 2008. Currently he is pursuing his Ph.D. in Artificial Intelligence. His research interest is the AI-based tools and techniques in Image processing particularly Steganography and Steganalysis. He is associated with the Department of Artificial Intelligence and Mathematical Sciences Sindh Madressatul Islam University Karachi.