2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1035-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCS53718.2022.9788309

Analysis of Uneven Stroke Prediction Dataset using Machine Learning

Atul Kumar Uttam
Department of Computer Engineering & Applications,
GLA University Mathura,
U.P., India
atul.uttam@gla.ac.in

***Abstract*— Stroke is a sudden interruption in the blood flow to a portion of the brain that can result in the loss of capacity to move certain parts of the body. In this work, all elements that impact this disease are studied, their influences are extracted using exploratory data analysis, and a model is constructed to forecast the disease's likely existence in a patient. Using under sampling and over sampling strategies, several data balancing techniques are used to increase model performance. As a consequence of the findings, a 98 percent accuracy rate has been achieved compared to earlier research. While the class distribution is severely skewed, there are relatively few examples of a class. This study reveals that over-sampling approaches yield better results than under-sampling ones. When sampling-based techniques are used instead of sampling-based approaches, the outcomes improve by 46%.**

Keywords*— Stroke Prediction, Machine Learning, Data Balancing, XGBoost.

I. INTRODUCTION

When it comes to fatalities and substantial disabilities, stroke is a major contributor to both. Stroke is the world's second-largest reason for mortality and the top reason for disability. It is becoming more common as the population ages. Furthermore, stroke affects a greater number of young individuals in poor and middle-income nations. It is the primary source of enduring impairment in the United States, particularly among older persons, who have the greatest rate of stroke. Of the 795,000 new stroke victims, 26% are still unable to do basic daily tasks, and 50% have restricted mobility owing to Hemiparesis. Other common causes of impairment include aphasia and depression. Disabled-adjusted life years (DALYs) were caused by stroke in developing and developed nations, with considerable geographical heterogeneity in disease burden in both. Stroke was the second greatest cause of DALYs worldwide and in developing nations. Because of modifiable risk factors, stroke can be prevented to a considerable extent. In high-income nations during the past two decades, a decrease in stroke incidence and DALYs (disability-adjusted-life-years) might be attributed to features such as smoking, obesity, blood pressure, diabetes as well as a lack of physical activity. This type of error is all too prevalent. There is a lack of funding for the finest medical facilities and workforce, imbalanced admittance to medical care, insufficient medical knowledge, and issues with devotion and observance that restrict the usefulness of chief and resultant anticipation in stroke treatment. Stroke incidence and prevalence are influenced by features such as age, gender, ethnicity, and socioeconomic status (SES). Several conditions can lead to cerebrovascular injuries or strokes. Stroke is one of the most deadly disorders for the elderly. As with a "heart attack," it damages the brain in the same way as it reparation the heart. It is the second foremost source of mortality in equally industrialized and developing countries alike. Even while strokes are extremely expensive and can cause long-term damage, they can also kill you. Up to 80% of stroke deaths can be averted by identifying or foretelling the beginning of a stroke in its earliest stage. A stroke occurs when the brain's bloodstream is cut off or reduced. Because the brain is deprived of oxygen and nutrients during a stroke, brain cells die. It's ideal to avoid this problem in the primary position by observing connected metabolic parameters. A prospective patient who relies only on clinical symptoms to determine whether or not more surgeries are necessary might be difficult for medical practitioners to assess. Most medical professionals refer to the brain and spinal cord injury produced by indiscretion in blood flow as a "stroke." Stroke can be viewed from many different perspectives, but on the whole, it elicits strong emotional reactions.

The precise prediction of stroke outcomes based on a set of predictive characteristics can aid in the identification of high-risk individuals and direct curative procedures, resulting in lesser morbidity. Identifying and verifying predictive variables may be done using a variety of models. When dealing with vast amounts of data from a variety of institutions, machine learning approaches provide an alternative that has the additional benefit of speedily putting together newly available data to augment forecast accuracy.

Both therapeutic and prognostic choices may be made more effective with early prediction in stroke patients. Many prognostic scoring systems have been developed as a result of this. According to recent breakthroughs in machine learning, medical applications of this technology have been extremely successful. Several studies [11, 12, 13] have demonstrated that machine learning algorithms are superior at portraying the dynamic and unpredictable character of human physiology.

II. LITERATURE SURVEY

The application of machine learning models can reliably predict long-term outcomes in acute stroke patients, according to a study by researchers [3]. For simplicity, prognostic ratings based on statistical studies employ only a few key factors, with their coefficients substantially rounded. Many factors, however, influence stroke outcomes, and these variables may have an impact on prediction, even if it is little. Their research revealed that the convolution neural network model outperformed the other models. The deep neural network model itself may be better suited to outcome prediction. Multiple layers of a complicated network may be useful in portraying the complexities of stroke patient outcomes. The theoretical underpinnings of the enhanced performance, however, are unknown.

The data from the National Population Health Science Data Warehouse platform were used to develop the risk forecast representation for stroke illness, according to the research article [4]. Other variables in the model included demographic data, lifestyle factors, checkup history, and a substantial test guide. Stroke risk was examined about several contributing factors. The Chi-square and ANOVA test were employed to analyze the combined data. A single-factor analysis was used to choose the influential elements, and the concluding features were resolute based on the findings of preceding studies. Males were shown to have a higher occurrence of the illness than females, which increased with increasing age. Smoking for an extended period was a significant hazard issue for stroke, as was the prevalence of hypertension with physiological symptoms and other associated features in those over the age of 65, according to the study.

Study [5] intended to construct a machine learning-based model that can forecast stroke in persons, with good performance that is suffering from symptoms of or is exposed to risks of stroke. In their paper [5], an approach based on machine learning and oversampling were developed. SMOTE, which employed a Random Forest classifier, exceeded earlier research in terms of performance.

According to the authors of work [6], an integrated machine learning approach combines data accusation, attribute assortment, and forecast. For both binary stroke forecast and predicting stroke danger, machine learning methods were shown to be superior to the Cox proportional hazards model, according to this study. When compared to other heuristics, the authors' conformist mean heuristic for attribute selection produces the best results. Their Margin-based Censored Regression, a new prediction approach, outperformed the Cox model in terms of concordance index.

To predict unexpected heart strokes, this study used input data such as hypertension, age, and average glucose levels as well as past BMI, heart diseases, and smoking status. The prediction was made using a variety of Support Vector Machine techniques. Ischemic stroke prognosis was recently examined by Cheng et al. [7]. They used 82 ischemic stroke serene data, two artificial neural network models, and 79 percent and 95 percent accuracy in their investigation. A study on stroke prediction using artificial intelligence was conducted by Singh et al. [10]. They utilized a new approach for predicting stroke in their analysis. They also applied the decision tree approach to principal component analysis after extracting features. The model was constructed using a CNN classification technique, and it was 97% accurate.

III. DATASET AND PREPROCESSING

We got our data from Kaggle [9] and utilized it in our investigation. Eleven separate variables are included in the dataset; they are age; gender; id; heart disease; hypertension; employment type; ever-married status; smoking status; residence type; glucose level; BMI; and one dependent variable; stroke. The dataset has a total of 5110 records out of which 4861 records of no stroke and 249 records of stroke are present. It shows that data is skewed, so first, we have to balance the dataset before applying a machine learning model. It is a classification problem as dependent feature stroke has two possible values 0 for no stroke and 1 for stroke. We have both numerical type as well as categorical type variables. Since the model takes numbers as input, we will have to encode the categorical features later on in the pre-processing step before feeding them to the model.

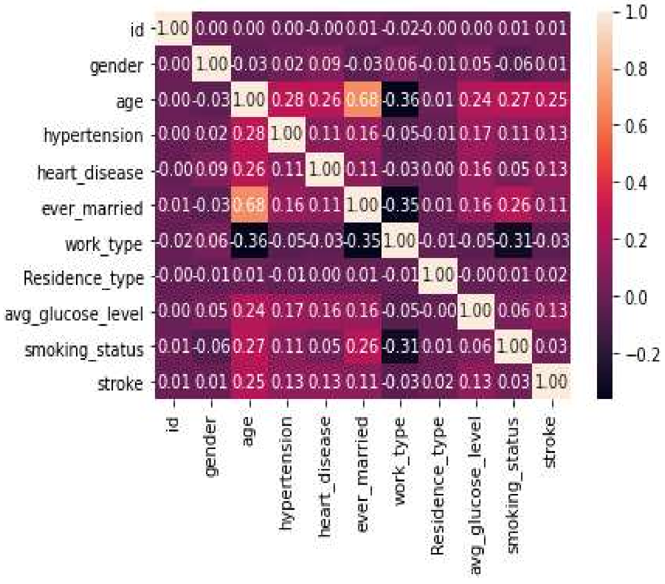

Fig. 1. Correlation between different features.

We have dropped the BMI feature as it has 4% missing values and it has no correlation with the dependent feature. After removing BMI feature from the study we have analyzed the correlation of different features with respect to target feature which is shown in figure 1. In this dataset gender, ever_married, work_type, Residence_type, and smoking_status are categorical features and the rest of the features are numerical, so we have encoded all the categorical features into numerical features. We see that in every feature, there are higher samples of no stroke (stroke=0) as compared to the other class. Hence it is a Highly Imbalanced dataset. Other categories in 'gender' can be ignored. Only 2 children have strokes and both are female. Older females in govt_jobs have a higher risk of stroke. The majority of people who had a stroke were working in the 'Private' sector.

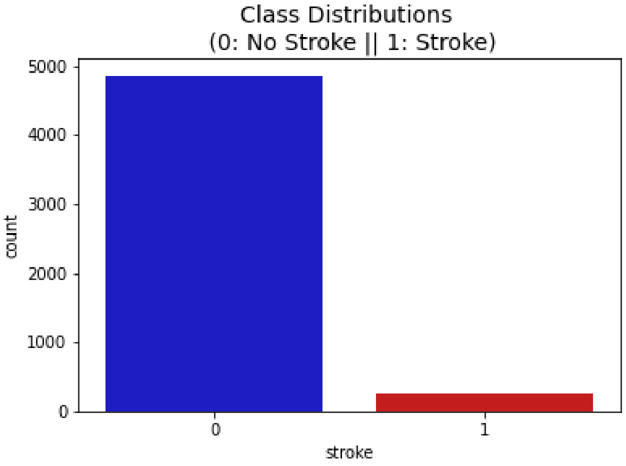

Fig. 2. Number of stroke(1) vs no stroke(0) records.

The descriptive statistics demonstrate characteristics of each variable for both groups: stroke and non-stroke patients. From the statistical analysis, we notice that mean values for all variables besides rural_res are significantly higher for stroke patients than non-stroke patients; therefore we can preliminarily conclude that individuals who are older with hypertension and/or heart disease with a higher level of glucose and BMI are in a risk group for stroke.

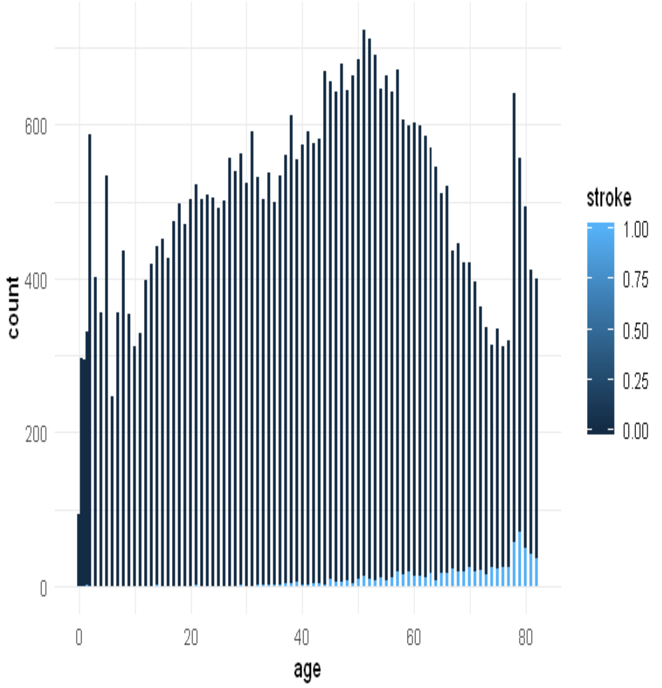

Fig. 3. Analysis of age on stroke.

Count of stroke and non-stroke patients by age. This illustration shows that the peak of stroke patients is in the age range 78-82. The stroke is not observed in the young group patients 0-32 years old. The average age for a stroke is about 68 years old. We observed that married people are more likely (by about 2%) to get a stroke than the ones who are not married. From the data analysis, we found that 1.5% out of people without hypertension typically get a stroke, while about 5% out of patients with hypertension get a stroke. The latter indicates that people with hypertension have a higher risk to get a stroke than people who don't have hypertension. People with heart disease have a higher chance to get a stroke. People who never worked have zero chance to get a stroke. However, self-employed have the highest risk factor 3.7%. It might be explained that they experience more stress and more responsibilities.

IV. Proposed Method

XGBoost, which stands for eXtreme Gradient Boosting, was used in this investigation. It's a decision tree boosting approach based on gradient boosted decision trees. One of the distinctions between XGBoost and gradient boosting is that it uses a superior regularisation strategy to reduce over-fitting. The open-source package 'XGBoost' provides machine learning algorithms based on gradient boosting approaches. The XGBoost is a program that boosts the performance of your computer. XGBClassifier is a classification class that works with the scikit-learn API.

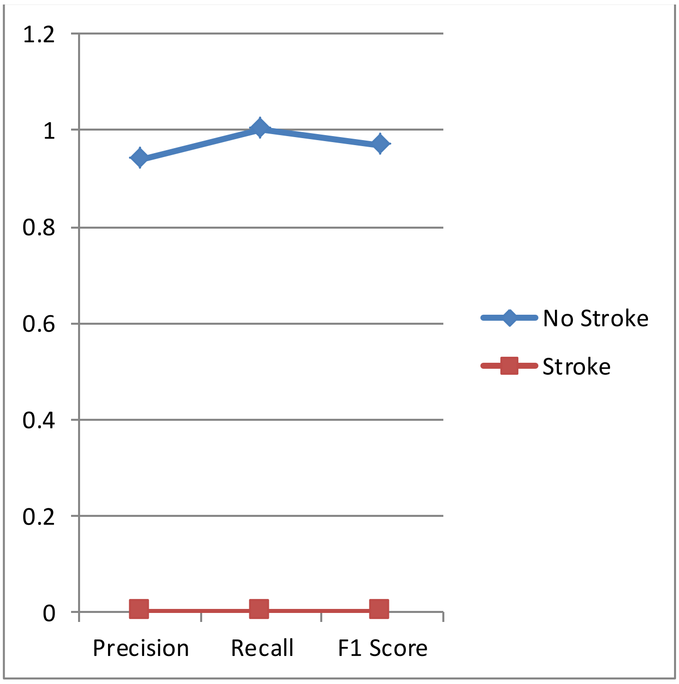

Fig. 4. Model performance without data balancing

From figure 4 it has been observed that the model fails to provide any significant result for stroke class, while it provides an accuracy of 94%. So to obtain a better and more realistic performance we have applied different data balancing techniques like Under Sampling (near miss, random under sampler) and Oversampling (auto, SMOTE minority, SMOTETomek_9). On applying above mentioned data balancing techniques the number of records in both classes has been balanced.

Table 1. In each target class the number of records after using data balancing strategies.

Techniques	Number of Records	
	No Stroke	Stroke
Original data	4860	249
Under Sampling-Near Miss	202	196
Under Sampling-Random Under Sampler	202	196
Over Sampling Auto	3929	3847
SMOTE Minority	3929	3847
SMOTETomek_9	3622	3586

After balancing the data we have again applied the same XGBoost model without changing any parameters and found significant improvement in the results.

Table 2. Performance score of the model using different data balancing techniques (0 No Stroke, 1 Stroke).

Technique	Class	Precision	Recall	F1 Score	Accuracy
Under Sampling Near Miss	0	0.65	0.64	0.65	0.67
	1	0.69	0.7	0.69	
Under Sampling Random Under Sampler	0	0.72	0.64	0.7	0.74
	1	0.72	0.83	0.77	
Over Sampling Auto	0	1	0.97	0.98	0.98
	1	0.97	1	0.98	
SMOTE Minority	0	0.95	0.97	0.96	0.96
	1	0.97	0.95	0.96	
SMOTETomek_9	0	0.95	0.97	0.97	0.97
	1	0.97	0.96	0.97	

Oversampling approaches perform better than under-sampling strategies, according to the table 2 data. And the auto approach is the most accurate oversampling strategy for the machine learning model available today.

V. RESULT AND DISCUSSION

After applying the supervised machine learning-based classification model (XGBoost) on a balanced dataset we have improved the accuracy and found that oversampling-based methods perform better than under sampling-based data balancing methods.

The main reason behind this is because of discarding of the significant amount of records by under-sampling techniques as support goes as low as 47 for no stroke and 53 for stroke class, which is enormously low data for the machine learning model. The Oversampling methods have higher support (for stroke 1013 and no stroke 931) which improves the learning ability of the machine learning model up to 46.27%.

Our analysis identified the factors which influence the probability of getting a stroke. They are high glucose level, older age, obesity, high blood pressure, heart disease, high-stress level (self-employed). By analyzing figure 5 and figure 6 we can observe that the proposed model provides 98% accuracy which is quite significant. Also model correctly identifies the true positive and true negative. People with age 65-85 have a high chance of getting a stroke. We can see (figure 4) how poor the model performs without training it with enough samples from both classes.

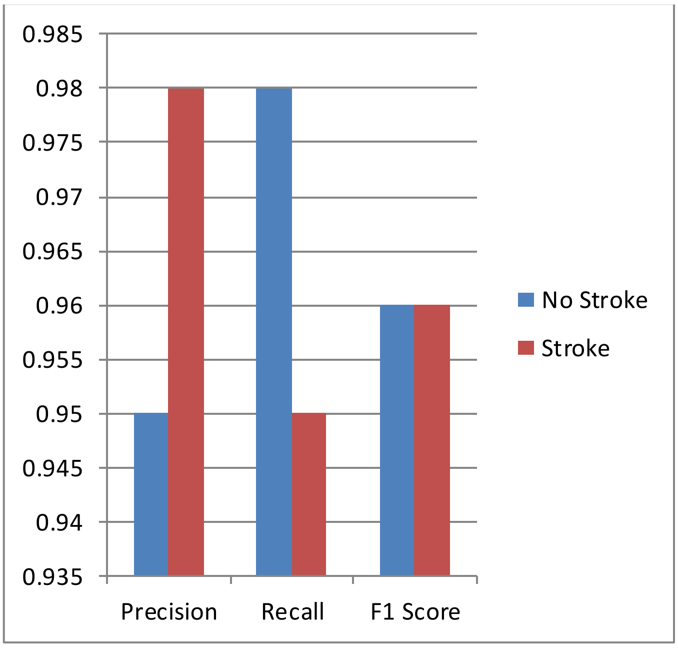

Fig. 5. Results of train and validation set by the proposed model.

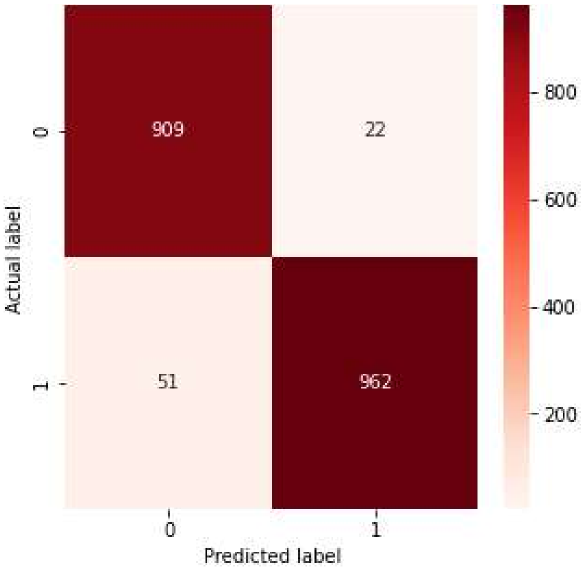

Fig. 6. Confusion Matrix of the proposed XGBoost model.

Before and after utilizing data balancing strategies in each class. Classification models may be evaluated by plotting their performance on the receiver operating characteristic curve (ROC curve). On this graph, we can see the True Positive and False Positive Rates. False Positives and True Positives grow when the classification threshold is reduced, resulting in more objects being categorized as positive. Figure 7 is an example of a typical ROC curve for the model under consideration.

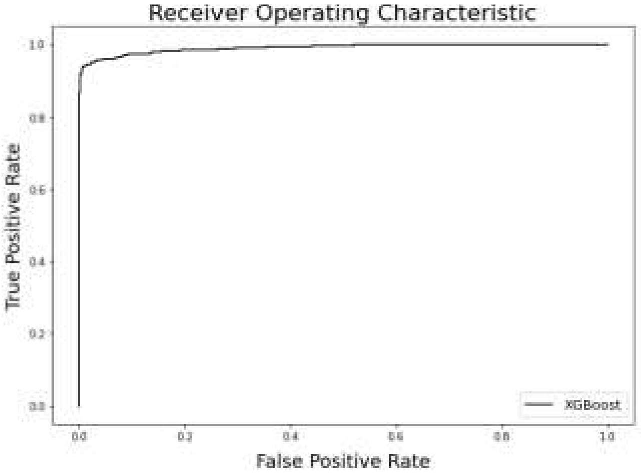

Fig. 7. ROC curve for proposed XGBoost model.

VI. Conclusion

Our analysis identified the factors which influence the probability of getting a stroke. They are high glucose level, older age, obesity, high blood pressure, heart disease, high-stress level (self-employed). In this work, we will try to analyze all the factors influencing this disease, extract their influence with exploratory data analysis and also build a model to predict the possible occurrence of disease in a patient. We have applied different data balancing techniques to improve the model performance using under-sampling and over-sampling methods. From the results, we have obtained better accuracy 98% from previous studies. This study also shows that over-sampling methods provide better results than under-sampling methods while the class distribution is highly skewed and we have very few instances of a class. The results have been improved by 46 % while we applied over-sampling-based methods.

References

[1] Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med. 2019 Nov;101:101723. doi: 10.1016/j.artmed.2019.101723. Epub 2019 Oct 23. PMID: 31813482.

[2] Katan M, Luft, “A. Global Burden of Stroke”. Semin Neurol. 2018 Apr;38(2):208-211. doi: 10.1055/s-0038-1649503. Epub 2018 May 23. PMID: 29791947.

[3] Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH, “Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. Stroke”. 2019 May;50(5):1263-1265. doi: 10.1161/STROKEAHA.118.024293. PMID: 30890116.

[4] Y. Liu, B. Ma and Y. Wang, "Study on prediction model of stroke risk based on decision tree and regression model," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4798-4801, doi: 10.1109/BigData52589.2021.9671409.

[5] Ferdib-Al-Islam and M. Ghosh, "An Enhanced Stroke Prediction Scheme Using SMOTE and Machine Learning Techniques," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579648.

[6] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee. 2010. An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, 183–192. DOI:https://doi.org/10.1145/1835804.1835830

[7] H. Puri, J. Chaudhary, K. R. Raghavendra, R. Mantri and K. Bingi, "Prediction of Heart Stroke Using Support Vector Machine Algorithm," 2021 8th International Conference on Smart Computing and Communications (ICSCC), 2021, pp. 21-26, doi: 10.1109/ICSCC51209.2021.9528241.

[8] Cheng CA, Lin YC, Chiu HW. Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. Stud Health Technol Inform. 2014;202:115-8. PMID: 25000029.

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

[10] M. S. Singh and P. Choudhary, “Stroke prediction using artificial intelligence,” in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, Aug. 2017, pp. 158–161.

[11] V. Jain and A. Yadav, "Analysis of Performance of Machine Learning Algorithms in Detection of Flowers," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 706-709, doi: 10.1109/ICICV50876.2021.9388599.

[12] Yadav A., Jain V., Kumar A. (2021) Performance Analysis of Machine Learning Algorithms in Credit Card Fraud Detection. In: Ranganathan G., Chen J., Rocha Á. (eds) Inventive Communication and Computational Technologies. Lecture Notes in Networks and Systems, vol 145. Springer, Singapore. https://doi.org/10.1007/978-981-15-7345-3_26.

[13] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1-3, doi: 10.1109/ICECA49313.2020.9297411.