WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Assessing the Health Risk of Different Factors Influencing Stroke of a Patient using Machine Learning

***Jyotirmoy Ghose[a], Md. Tasnim Alam[b],Md. Iftekhar Hossian[c], Dipok Chandra[d]***

[a] Lecturer, North Bengal International University, Chowddopai, Natore Road, Motihar, Rajshahi, Bangladesh.
[b] Lecturer, North Bengal International University, Chowddopai, Natore Road, Motihar, Rajshahi, Bangladesh
[c] Lecturer, North Bengal International University, Chowddopai, Natore Road, Motihar, Rajshahi, Bangladesh
[d] Student, North Bengal International University, Affiliated by Jyotirmoy Ghose, Department of Computer Science and Engineering,Chowddopai, Natore Road, Motihar, Rajshahi, Bangladesh
DOI: https://doi.org/10.55248/gengpi.2023.4235

**A B S T R A C T**

A stroke is a disease in which the blood arteries in the brain burst, harming the brain. If the brain's supply of blood and other nutrients is compromised, symptoms could develop. The World Health Organization states that stroke is the primary cause of death and disability worldwide (WHO). The severity of a stroke can be lowered by being aware of the many stroke warning signs as soon as possible. There are numerous machine learning (ML) models that have been developed to predict the likelihood of a brain stroke occurring. The technique was created using the freely available Stroke Prediction dataset. The accuracy percentage of the models used in this investigation is noticeably higher than in past studies, demonstrating the models' higher reliability

**Keywords:** Ischemic Stroke, Hemorrhagic Stroke, TIA, k-NN, Confusion Matrix, Decision Tree, Random Forest, Machine Learning, ECG, Glucose

## 1. INTRODUCTION

A stroke is a condition where there is insufficient blood supply to the brain, which results in cell death. Strokes can be ischemic or hemorrhagic, which refers to a lack of blood flow or bleeding, respectively. Both stop certain brain functions in their tracks. The inability to move or feel on one side of the body, difficulties understanding or speaking, dizziness, or loss of vision to one side are all potential signs and symptoms of a stroke. After a stroke, symptoms and signs frequently start to show very quickly. Transient ischemic attack (TIA), often known as a mini-stroke, is the type of stroke that occurs when symptoms last less than one or two hours. A severe headache could potentially be a symptom of a hemorrhagic stroke.[1] A stroke may leave behind permanent symptoms. Pneumonia and a loss of bladder control are examples of long-term consequences. Blood pressure is the biggest risk factor for stroke. Other risk factors include atrial fibrillation, high blood pressure, smoking, obesity, diabetes, a prior TIA, and end-stage kidney disease. Although there are other, less frequent causes of ischemic strokes, blood vessel blockage is frequently the culprit. Either bleeding into the brain itself or into the space between the membranes of the brain might result in a hemorrhagic stroke. An aneurysm in the brain that has ruptured may cause bleeding. A physical examination is often used to make a diagnosis, which is then confirmed by imaging tests like a CT or MRI. Although ischemia, which early on often does not show up on a CT scan, can be ruled out by a CT scan, bleeding can still be ruled out by the scan. To identify risk factors and rule out other potential causes, further procedures such blood tests and electrocardiograms (ECGs) are performed. [2] The symptoms of low blood sugar may be similar. Reducing risk factors, brain artery surgery for those with significant carotid narrowing, and warfarin for those with atrial fibrillation are all methods of prevention. Doctors might advise statins or aspirin for prophylaxis. TIAs and strokes frequently necessitate emergency care. A medicine that can dissolve the clot might be used to treat an ischemic stroke if it is identified within three to four and a half hours. Surgery is beneficial in some hemorrhagic stroke cases. Stroke rehabilitation is a type of treatment used to try to regain lost function; it is preferable to receive this kind of care in a stroke unit, but these facilities are rare in many parts of the world.[3].

## 2. METHODOLOGY

### *2.1 Machine Learning (ML):*

Machine learning is an artificial intelligence subfield that is defined as a computer's ability to imitate intelligent human behavior. Artificial intelligence (AI) systems are used to solve complex problems in a way that is similar to how people solve this problem. The study of complicated algorithms that may improve themselves autonomously based on personal experience and data is known as machine learning (ML)

### *2.2 Data collection in ML:*

Data is any discrete fact, measurement, or piece of information that can be quantified analytically. A more complex definition of information is a collection of subjective or numerical data. the single value of a single variable, whereas a datum (unique piece of information) is attributes relating to at least one person or object. Although the phrases "information" and "data" are sometimes used synonymously, they have different connotations in this context. Data can be obtained from various data sources such as APIs, File, and Database

### *2.3 CSV Format:*

A structured text file called a "comma-separated values" (CSV) file contains values that are separated by commas. Each line of the sentence alludes to a different piece of information. One or more fields make up each record, and they are all separated by commas.The name of this method was inspired by the usage of commas as field separators

### *2.4 Python Libraries used in Data Analysis:*

Python is a programming language used to create websites and software applications, automate processes, and conduct data analysis. Python is a strong programming language that isn't specialized in solving any particular issue, therefore it may be used to construct a wide range of applications.

### *2.5 Exploratory Data Analysis (EDA):*

One of the most efficient data science techniques today is exploratory data analysis. The difference between data processing and data exploration is frequently unclear to people who are just starting out in data science. The terms aren't really distinct, yet they have different purposes. To find trends in a data set, data analysis uses statistics and probability.

### *2.6 Confusion Matrix in ML:*

The performance of the classification models for a certain set of test data is evaluated using a matrix called the confusion matrix. Only after the true values of the test data are known can it be determined.

### *2.7 Machine Learning Classifiers models:*

A classifier is a kind of Machine Learning procedure utilized in information science to dole out a class name to include information. A picture acknowledgment classifier, for instance, can mark an image (e.g., "vehicle," "truck," or "human"). Classifier calculations utilize progressed numerical and measurable ways to deal with gauging the opportunity of an information input being arranged with a particular goal in mind. In the picture acknowledgment model, the classifier numerically predicts whether an image is probably going to be a vehicle, a truck, or something different. Here are a few classifiers utilized in the exploration gave underneath

**2.7.1** **Decision Tree:** A decision tree is a judgment tool that uses a tree-like model of options and possible consequences, such as chance events outcomes, cost objects, and utility, to make judgments.

**2.7.2** **Logistic Regression:** Supervised learning is demonstrated using logistic regression. It is used to compute or forecast the likelihood of a binary (yes/no) event occurring. A logistic regression example would be using machine learning to assess whether or not a person is likely to be infected with stroke.

**2.7.3** **Random Forest:** is a Supervised Machine Learning Algorithm as often as possible used in Classification and Regression applications. It develops choice trees from a few examples and utilizations their unmistakable greater part for classification and normal for relapse.

**2.7.4** **k-Nearest Neighbour (KNN):** The k-Nearest Neighbour strategy is a lethargic learning procedure that saves all events in n-layered space that match to preparing of important informative elements

## 3. RESULT AND DISCUSSION

The Results (or Findings) section follows the Methods and precedes the Discussion section. This is where the authors provide the data collected during their study. That data can sometimes be difficult to understand because it is often quite technical. Do not let this intimidate you; you will discover the significance of the results next.

I found some different results upon my analysis


Fig 3.1.1: In the dataset there are more female candidates than male.

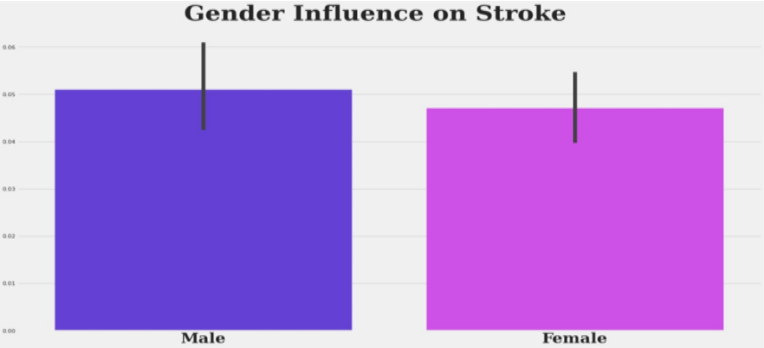


Fig. 3.1.2 - (a) Male; (b) Female: Male candidates are usually affected by stroke than female candidates

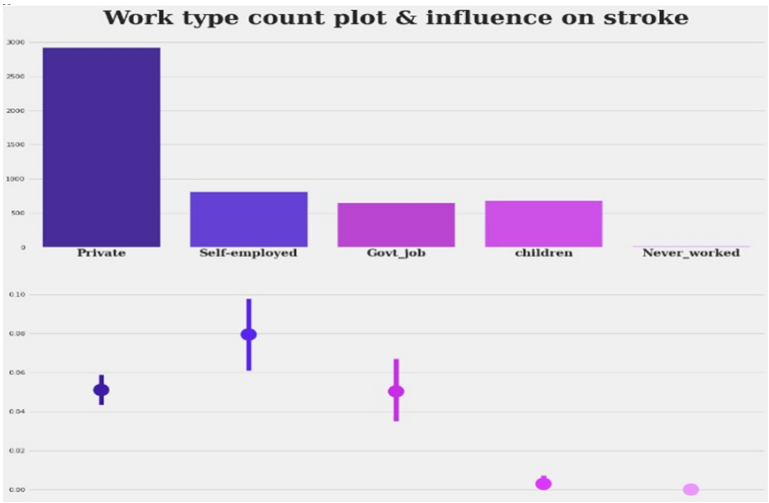


Fig 3.1.3: Private Job holders are in big numbers but Self-employed candidates are usually affected by stroke than others

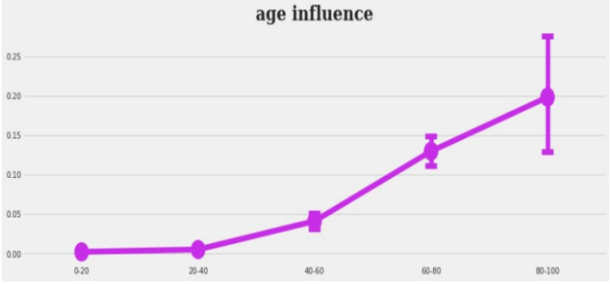


Fig 3.1.4: Older persons are more suffered from the risk of getting stroke

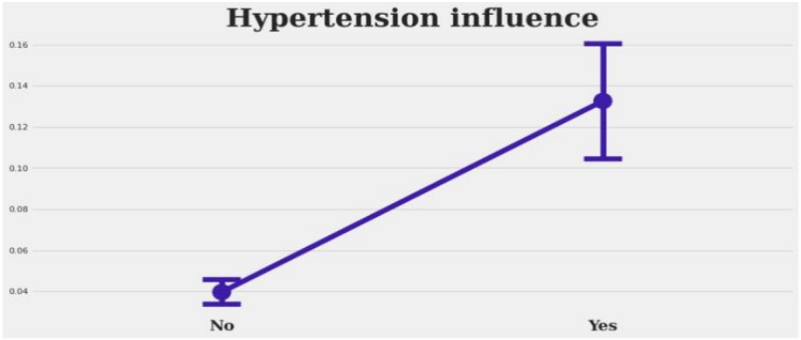


Fig 3.1.5: Hypertension patient are more suffered from the risk of getting stroke

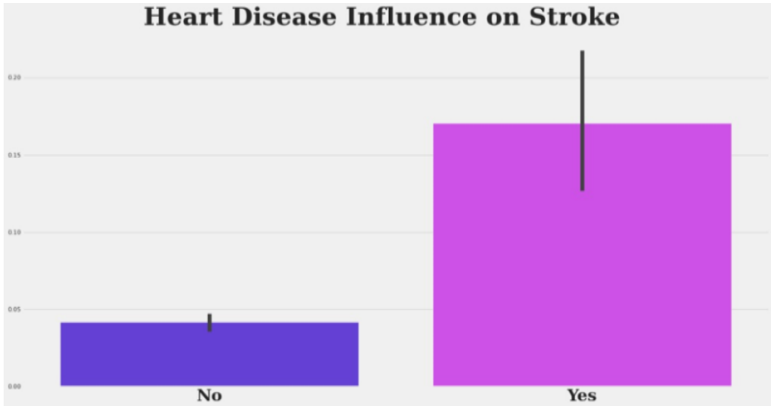


Fig 3.1.6: Heart disease patients are more suffered from the risk of getting Stroke

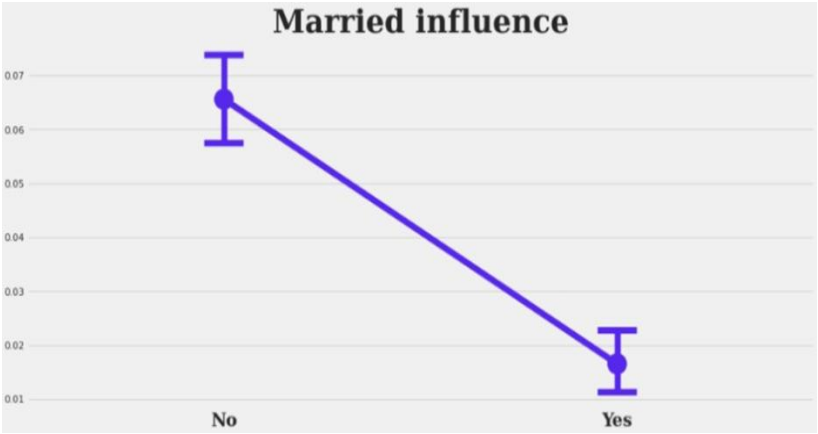


Fig 3.1.7: Non married persons are more suffered from the risk of getting stroke

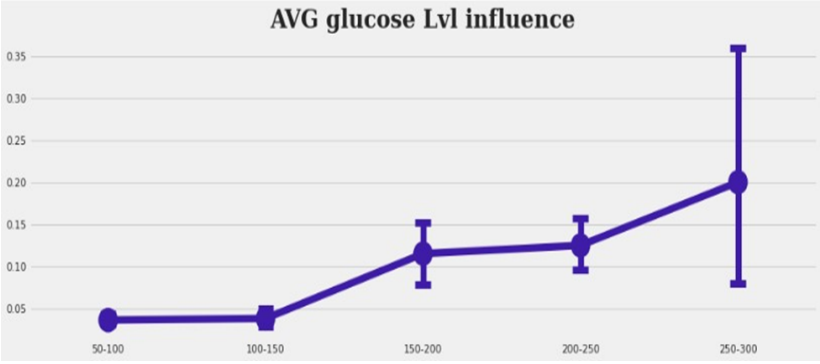


Fig 3.1.8: Persons with body mass 40-60 are more suffered from the risk of getting stroke

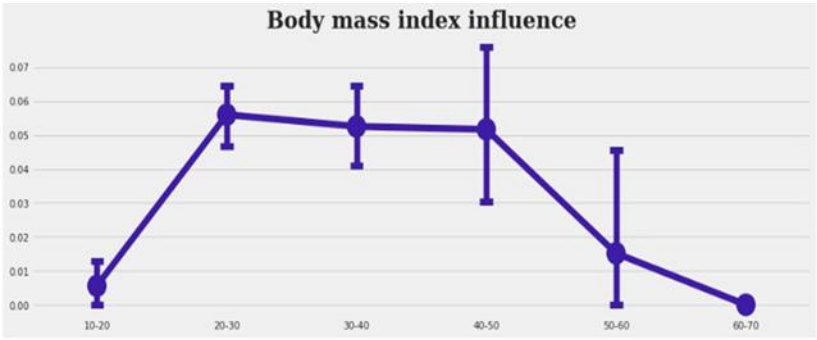


Fig 3.1.9: Persons with body mass 40-60 are more suffered from the risk of getting stroke

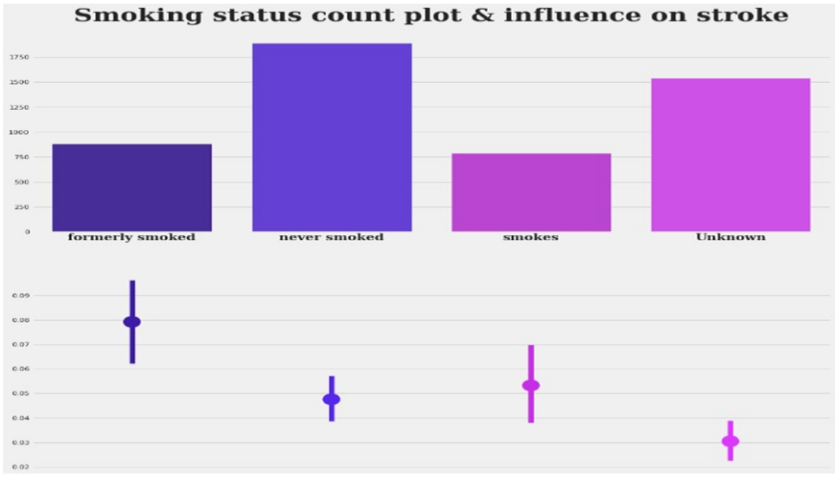


Fig 3.1.10: Persons who are formerly smoked and regularly smokes are more suffered from the risk of getting stroke

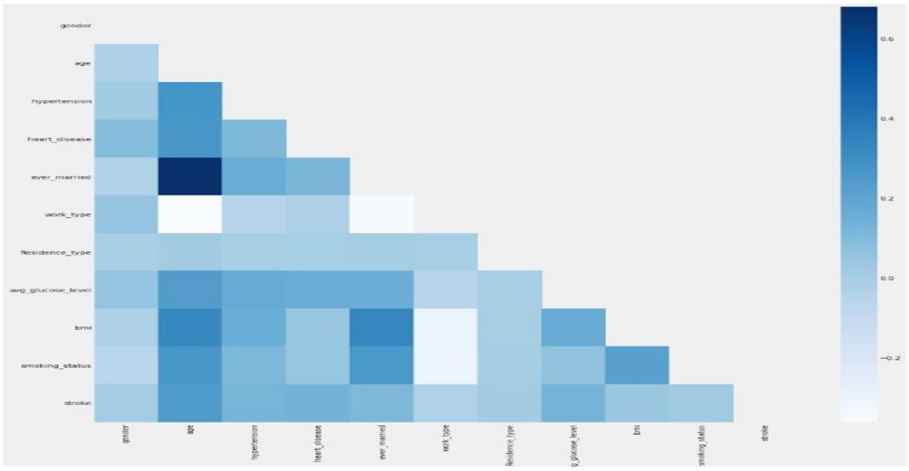


Fig 3.1.11: By correlation matrix I found that old aged persons who are married and regularly getting smoked are mainly at risk of getting stroke in the long process.

### *3.2 ACCURACY TEST*

Here a confusion matrix and many classifiers for the accuracy test has been utilized. The confusion matrix of Random Forest that been measured is shown below.

True Positive = 1336, True Negative = 1413, False Positive = 110, False Negative = 57

For Model RF the result has been shown bellow

Score = 0.943, Precision = 0.92, Recall = 0.961, F1 = 0.944

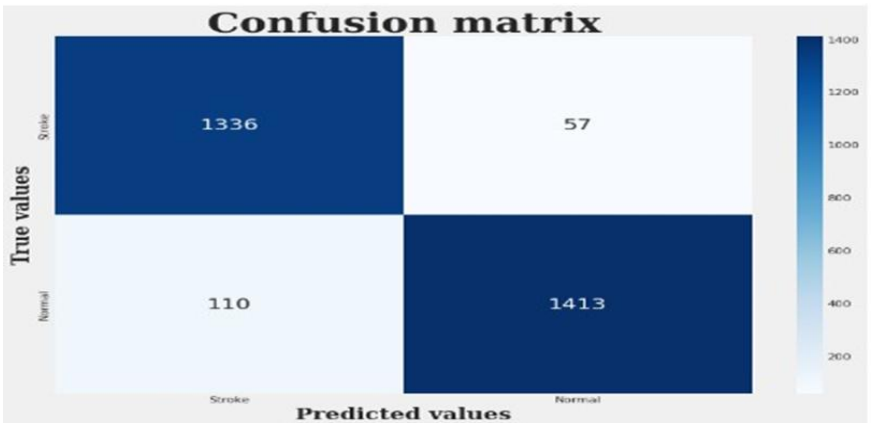


Fig 3.2.1: Confusion Matrix for Random Forest.

Measured value of confusion matrix of KNN is given below,

True Positive = 1226, True Negative = 1431, False Positive = 220, False Negative = 39

For Model kNN

Score = 0.911, Precision = 0.867, Recall = 0.973, F1 = 0.917

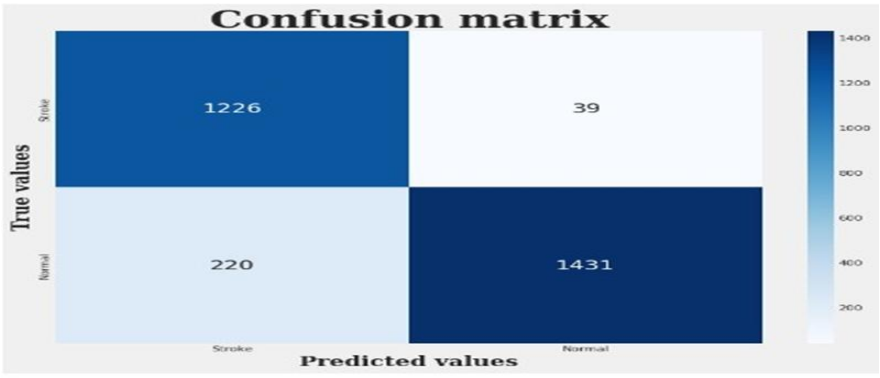


Fig 3.2.2: Confusion Matrix For KNN Classifier.

Measured value of confusion matrix of Gradient Boosting is given below,

True Positive = 1359, True Negative = 1403, False Positive = 90, False Negative = 72

For Model GradBoost

Score = 0.947, Precision = 0.942, Recall = 0.954, F1 = 0.948

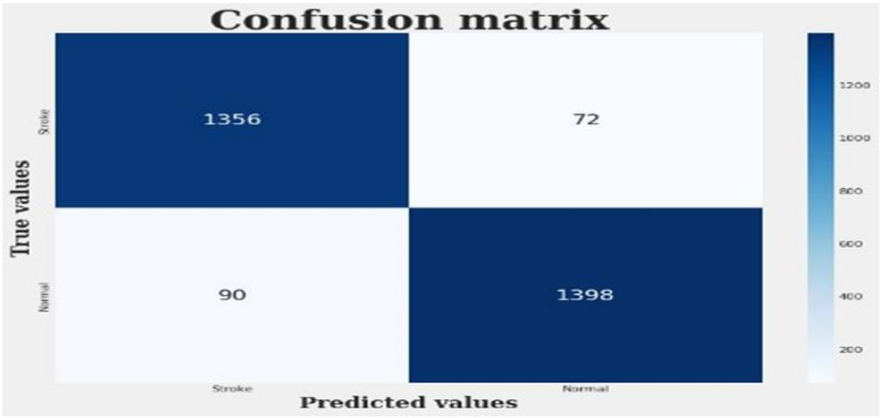


Fig 3.2.3: Confusion Matrix for Gradient Boosting

Measured Neural Network score = 0.928

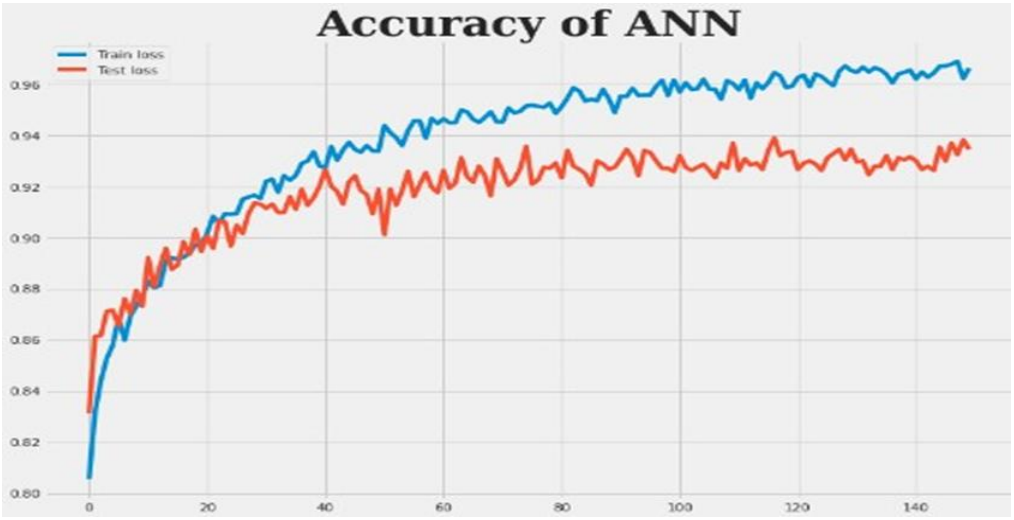


Fig 3.2.4: Test loss is bigger than trained loss in ANN

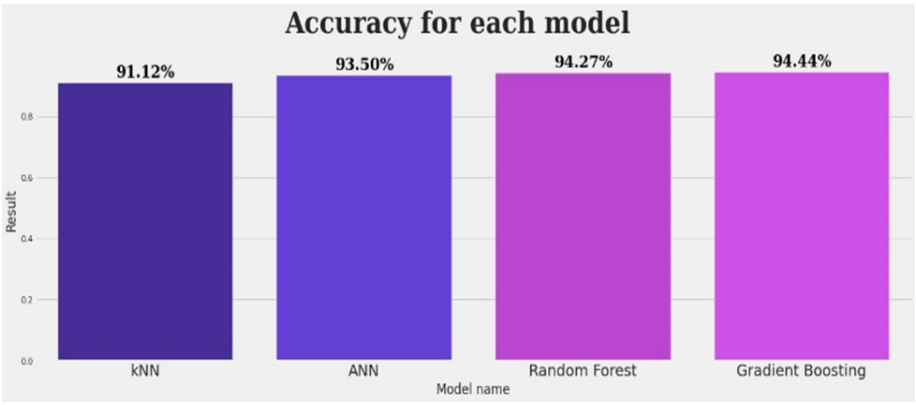


Fig 4.16: Accuracy detection

## 4. CONCLUSION:

When blood flow to a portion of the brain is disrupted due to a damaged or clogged blood artery, a stroke ensues. Hemorrhagic or ischemic strokes are both possible. When a blood vessel in the brain ruptures or breaks, blood leaks into the brain, causing a hemorrhagic stroke. When a blood channel transporting blood to the brain is stopped or restricted by severely narrowed arteries or a blood clot, an ischemic stroke develops. In this work, I analyze all the factors influencing this disease (STROKE) to extract their influence with exploratory data analysis and classifier models. In the dataset there are more female candidates than male. Basically, my work is based on the difference between male female candidates of how stroke can take effect on them. Private job holders are in big numbers but Self-employed candidates are usually affected by stroke than others. Older persons are more suffered from the risk of getting stroke. Hypertension patients are more suffered from the risk of getting stroke. Surprisingly, Unmarried persons are more suffered from the risk of getting stroke. Glucose level of 250-300 is more suffered from the risk of getting stroke. Persons with body mass 40-60 are more suffered from the risk of getting stroke. Persons who are formerly smoked and regularly smokes are more suffered from the risk of getting stroke. I take the accuracy tests also for successful project work. In Here I found that is the

**References**

*[1]https://www.nhlbi.nih.gov/health/stroke Retrieved 2023-01-14.*

*[2] Hu, A., Niu, J. and Winkelmayer, W.C., 2018, November. Oral anticoagulation in patients with end-stage kidney disease on dialysis and atrial fibrillation. In Seminars in nephrology (Vol. 38, No. 6, pp. 618-628). WB Saunders.*

*[3] Emergency Neurology edited by Karen L. Roos page-355*

*[4] Fatahzadeh, M. and Glick, M., 2006. Stroke: epidemiology, classification, risk factors, complications, diagnosis, prevention, and medical and dental management. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, 102(2), pp.180-191.*

*[5] Gratton, C. and Jones, I., 2010. Research methods for sports studies. Routledge.*

*[6] Zimmerman, A., 2007. Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. International Journal on Digital Libraries, 7(1), pp.5-16.*

*[7] Berry, M.W., 2010. SPARSE TENSORS DECOMPOSITION SOFTWARE.*

*[8] McKinney, W., 2012. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".*

*[9] Zardari, M.A., Jung, L.T. and Zakaria, N., 2014, June. K-NN classifier for data confidentiality in cloud computing. In 2014 International Conference on Computer and Information Sciences (ICCOINS) (pp. 1-6). ieee.*

*[10] Gupta, P. and Sehgal, N.K., 2021. Machine learning algorithms. In Introduction to Machine Learning in the Cloud with Python (pp. 23-77). Springer, Cham.*

*[11] Starzacher, A. and Rinner, B., 2008, December. Evaluating KNN, LDA and QDA classification for embedded online feature fusion. In 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing (pp. 85-90). IEEE.*

*[12] Dev, V.A. and Eden, M.R., 2019. Formation lithology classification using scalable gradient boosted decision trees. Computers & Chemical Engineering, 128, pp.392-404.*

*[13] Bakshi, K. and Bakshi, K., 2018, March. Considerations for artificial intelligence and machine learning: Approaches and use cases. In 2018 IEEE Aerospace Conference (pp. 1-9). IEEE.*

*[14] Tschang, F.T. and Almirall, E., 2021. Artificial intelligence as augmenting automation: Implications for employment. Academy of Management Perspectives, 35(4), pp.642-659.*

*[15] Young, J.B. and Forster, A., 1991. Methodology of a stroke rehabilitation trial. Clinical rehabilitation, 5(2), pp.127-133.*

*[16] Amann, J., 2021. Machine Learning in Stroke Medicine: Opportunities and Challenges for Risk Prediction and Prevention. Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues, pp.57-71.*

*[17] "Data vs Information - Difference and Comparison | Diffen". www.diffen.com. Retrieved 2023-01-14.*

*[18] "SuperCalc², spreadsheet package for IBM, CP/M". Retrieved December 11, 2017.*

*[19] Rossant, C., 2018. IPython Interactive Computing and Visualization Cookbook: Over 100 hands-on recipes to sharpen your skills in high-performance numerical computing and data science in the Jupyter Notebook. Packt Publishing Ltd.*

*[20] "Matplotlib Lead Developer Explains Why He Can't Fix the Docs—But You Can –NumFOCUS". NumFOCUS. 2017-10-05. Retrieved 2023-01-14.*

*[21] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. and Van Der Walt, S.J., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, 17(3), pp.261-272.*

*[22] Shukla, R., Gyanchandani, M., Sahu, R. and Jain, P., 2020. Analytical Approach to Genetics of Cancer Therapeutics through Machine Learning. In Soft Computing Applications and Techniques in Healthcare (pp. 1-10). CRC Press.*

*[23] Quinlan, J.R., 1987. Simplifying decision trees. International journal of man-machine studies, 27(3), pp.221-234.*

*[24] Mills, P., 2011. Efficient statistical classification of satellite measurements. International Journal of Remote Sensing, 32(21), pp.6109-6132.*

*[25] Ben-Hur, A., Horn, D., Siegelmann, H.T. and Vapnik, V., 2001. Support vector clustering. Journal of machine learning research, 2(Dec), pp.125-137.*