**RESEARCH ARTICLE** **SPECIAL ISSUE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S050

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

# Stroke prediction using 1DCNN with ANOVA

*Mallikarjunamallu K [1], Khasim Syed [1]*

*[1]School of Computer Science and Engineering VIT -AP University, Amaravati, Andhra Pradesh , 522237, India.*

Emails: mallikarjuna.21phd7147@vitap.ac.in, syed.khasim@vitap.ac.in

**Article History**

*Received: 28 February 2023*
*Accepted: 21 March 2023*
*Published: 28 May 2023*

***Keywords***:

*KN;*
*SVM;*
*LR;*
*RF;*
*GBS;*
*LGBM;*
*1DCNN*

**Abstract**

*Stroke and heart disease are among the most common outcomes of hypertension. Each year, heart disease, stroke, and other cardiovascular disorders claim the lives of more than 877,500 people in the United States, making them the first and fifth leading causes of death, so being able to predict them early helps save lives. A lot of research has been done to reach this goal. Machine learning models are mostly used for this purpose. For the first time in this study, we have used the Deep Learning (DL) model, i.e., one dimensional convolutional neural network (1D CNN) . In this study, first we extracted important features using the Analysis of variance (ANOVA) method. Then the data set with the new features that came up was given to the model. Then we compare all machine learning algorithms—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest Classifier (RF), Gradient Boosting Clas-sifier (XGB), and LoLight gradient boosting machine classifier (LGBM)—with 1DCNN. Recall, the F1 score, accuracy, and precision are some of the confusion metrics used to assess the effectiveness of the results.The results show that when used on reprocessed data, the proposed model performs best and is more than 98% accurate.*

## 1. Introduction

The fifth industrial revolution has begun. Health technology has advanced greatly. Modern medicine cannot cure all ail-ments.When brain cells don't get enough oxygen and nutrients, a stroke occurs. Treatment for a stroke is considered an emergency. The earlier a disease is detected and treated, the less damage there will be to the brain and other organs. According to the WHO, 15 million people suffer from stroke each year, and 4-5 minutes later, they die.There are two kinds of strokes: ischemic and hemorrhagic. Ischemic strokes happen when a clot blocks a blood vessel that brings blood to the brain. When a blood vessel bursts and lets blood into the brain, this is called a hemorrhagic stroke.Heart attacks and strokes happen when the heart doesn't get enough blood (Babu et al.). General identification of human brain and heart stroke is displayed in Fig. 1. Blocking oxygen or blood flow to the heart Muscle blockage that occurs during myocardial infarction Arteries that supply blood to the brain can cause heart attacks. flat Diseases vary, but their risk factors Their contributions are very similar. unhealthy food, smoking, diabetes, sedentary lifestyle, poor health Alcohol Use, Hypertension, and Family History is all

risk factors. heart attack detection and Seeing a doctor early is not only helpful Not only does it help you live longer, it also helps you avoid heart problems future. Machine learning has evolved The most difficult area of modern technology. this

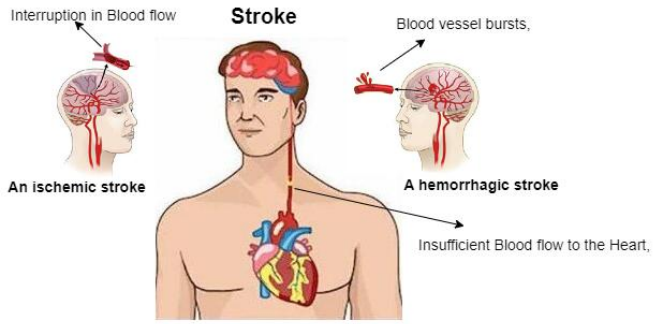



**FIGURE 1. Architecture of Stroke**

is the model can evaluate the artificial intelligence of the data, Discover patterns and expect little or no results interpersonal communication. Various machine learning techniques It can be used to predict heart attacks in humans. based on These input functions are A dataset is suggested for training the model Use a model to predict heart attacks for multiple people Various machine learning methods (M. Wang, Yao, Chen, et al.).

### *1.1. Description of the Problem*

The fifth economic revolution is taking place. Health tech-nology has improved greatly. Modern medicine cannot cure all diseases. Stroke is the worst modern disease. A brain blood vessel clot or burst limits oxygen and nutrients. Brain cells die, causing a stroke. It targets brain arteries and veins. Bangladesh's silent killer, stroke, affects 21.5% of all elderly people.Strokes caused 5.7 million deaths, 13% of global deaths, in 2016.The population's lack of medical professionals and infrastructure is a major issue. Thus, many studies have examined how computers could automate or assist healthcare workers in early disease detection and treatment (Jalajajayalakshmi, Geetha, Ijaz, et al.).

### *1.2. The Contribution of the Author*

Below is a list of the author's contributions.

- The pro-posed method was improved in terms of data consistency.
- In this study, we investigated the impact of DL and several different ML models on e a number of ML algorithms applied to the Framingham data set with the intention of predicting heart rate.
- I found that CNN has the best accuracy at 98.9%, and the F1 score is 98%. This means that the CNN's performance significantly outperforms the other machine learning models we investigated.

## 2. Literature Survey

The classification of heart attacks has been suggested in a great number of different ways, some of which are discussed in this section.

In (Beunza et al.), the author used SVMs to achieve 75% accuracy. Using a significance level (p-level) to pre-process the dataset (Framingham data). PCA, UCA, particle swarm optimization, and other feature selection methods can increase accuracy. In (Fisher, Smith, Walsh, et al.), The authors used data from 4395 patients to train an ANN (Artificial Neural Network) and then used data from 861 patients to test it. The test set scored 76.7% and the training set 74.3%. In (Mohan, Thirumalai, G. Srivastava, et al.), They looked at the UCI dataset to try to predict the risk of heart disease.The quality of the data is improving. HRFLM yielded 88.4% accuracy. Feature selection and a larger dataset can also improve accuracy. In (Patil, Shastry, Ashokumar, et al.), the author compares DT, NAVIE B, NN, and KNN. The accuracy of k-nearest neighbour was 68.2%. A MIXED ML model classified heart disease with 88.4% accuracy (Verma, S. Srivastava, Negi, et al.). Particle swarm optimization is used on the Indira Gandhi Medical College data. Using decision trees and different machine learning algorithms to compare three data sets gave an accuracy of 88.09%

(Babič et al.). In (Chaurasia, Pal, et al.), ID3-CART decision tree classifiers were used on UCI heart disease data. The CART classifier's accuracy reached 83.49%. K-fold cross-validation was used to prevent model bias. In (Anitha, Sridevi, et al.), the authors preprocessed missing values using the UCI heart disease repository. They achieved a maximum accuracy of 86.6% using SVM and k-NN machine learning algorithms. Optimized feature selection improves accuracy. The authors used UCI machine learning repository models in (Sharma et al., "Comprehensive Analysis of Feature Selection on Early Heart Strok Prediction"). Stochastic gradient descent helped them reach 87.69% accuracy. In ("Heart Stroke Risk Analysis: A Deep Learning Approach"), the author discussed a logistic regression-based approach for heart disease predic-

tion that outperformed other machine learning algorithms with 86.8% accuracy. The dataset was not specified. (Fang, Huang, Z. Wang, et al.)Gang Fang used publicly available data from the International Stroke Trial (IST) to predict what would happen 6 months after an ischemic stroke (IS). These systems were CNN, LSTM, and Resent, which achieved 0.83% accuracy. (Tougui, Jilbab, Mhamdi, et al.) compares five machine learning models and artificial neural networks. The artificial neural network had 85.86% accuracy. (Emon et al.)Minhaz Uddin Emon demonstrated how hypertension, body mass index, heart disease, average blood glucose level, smoking status, previous stroke, and age can be used to predict an early stroke. Ten classifiers were trained using these distinguishing features. The results of the baseline classifiers were weighted to maximize accuracy. The proposed studyreceived a validation score of 97%.

Depending on the above previous techniques, it can be said that the classification accuracy of the existing methods is low.because they don't take feature selection techniques and data sets into account properly. This is a problem that the proposed work tries to fix.

## 3. Proposed Methodology

This section gives an overview of the system's architecture and explains the suggested algorithm and other important ideas.

### *3.1. System Architecture*

A graphical illustration of how the proposed systems in-teract with one another is provided in this section as Fig. 2. It requires a number of procedures, such as data cleansing and feature engineering, separating the reprocessed data into training and testing.

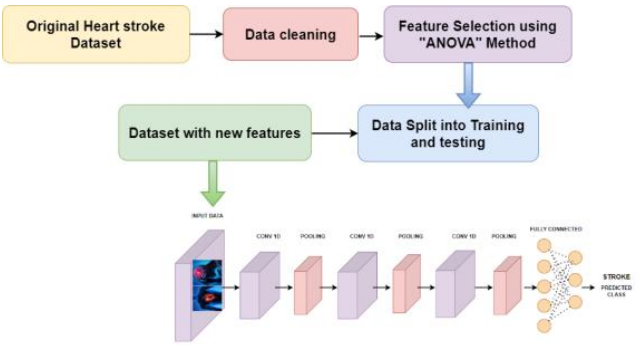



**FIGURE 2. Architecture of the proposed model**

This model we proposed is divided into two parts. First, we cleaned the original Stork data set. Then, using the wrapper method, we selected some important features and created a new data set. We fed that data set into the CNN model. In the second half of the model, the CNN comprises three one-dimensional CNN layers, and then three pooling layers are used as the last fully connected layer for stroke class prediction.

### *3.2. Details of the Dataset*

The data used in this study was obtained from a medical clinic in Bangladesh. It is the document containing 5110 people's information, and all of the features are described at this point:

Information regarding the stroke Data Set is shown in Table.I.

All of the stroke attributes make up the decision class in this case, whereas the other attributes make up the response class (Akter et al.).

**TABLE 1. Information regarding the stroke dataset**

| S.No | Feature | Description |
|---|---|---|
| 1 | age | This attribute means a person's age. Numerical data. |
| 2 | gender | This indicates gender. Categorical. |
| 3 | hypertension | This indicates hypertensive or not. Numbers. |
| 4 | work type | Person job scenario is this attribute. Categorical. |
| 5 | residence type | Person's living situation. Categorical data |
| 6 | heart disease | This indicates person has heart problems. Numbers. |
| 7. | avg glucose level | This indicates a person's glucose level. Numbers. |
| 8 | bmi | Person's body mass index. Numbers. |
| 9 | ever married | This indicates marital status. Categorical. |
| 10 | smoking Status | Smoking condition. Categorical. |
| 11 | stroke | This indicates a stroke incidence. Numbers. |

### 3.3. Pre-Processing of Data

#### 3.3.1. Cleaning and organising :

During this step, the database is analysed to look for any errors, missing values, or irrele-vant features. In this stage, we will handle null values, deal with outliers, encode category categories, and transform NaN values.(e.g. using modes for categorical variables, means for continuous variables), etc. The Framing ham data set has 540 missing values, filled with the mean and average.

#### 3.3.2. "ANOVA" Feature Extraction :

A common way to find a good treatment method is to look at how long it took patients to get better. We can show how much these three treatment samples differ from each other by comparing them using a statistical method. ANOVA is the name of the method used to compare samples based on their means. Analysis of variance (ANOVA) is a statistical method used to determine if the means of two or more groups are significantly different from each other. A ANOVA compares the means of different sam-ples to determine how one or more factors affect them. Using ANOVA, we can prove or disprove that all drug treatments are equally good . In this manuscript, we used the ANOVA method for feature selection. After applying this method, we obtained the 10 important features shown in Fig. 3.

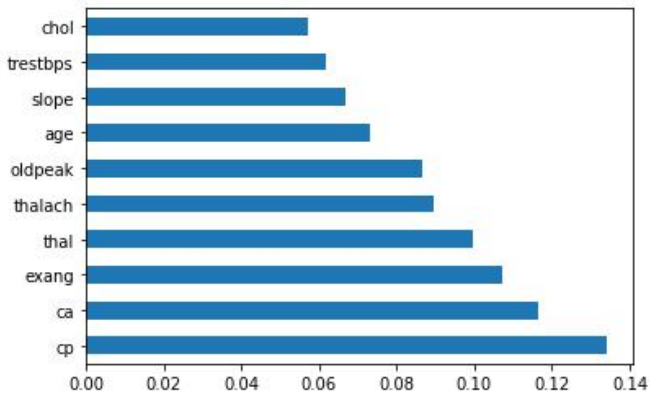



**FIGURE 3. Correlation between independent variables**

### 3.4. Creation of Models: Effectiveness of Prediction

In this step, the preprocessed data is first split into two groups, a training set and a test set. The training set contains 80% of the data and the test set contains 20%. Then use predictive models such as K-Nearest Neighbors, Support Vec-tor Machines, Logistic Regression, Random Forest Classifiers, Gradient Boosted Classifiers,LGBM and CNN. The model above is trained on the training set and then tested on the test set.

#### 3.4.1. K-Nearest Neighbour (KNN):

It is an algorithm for supervised machine learning. It is also known as a "delayed learning algorithm" because it does not require any additional steps for training. This algorithm is effective because it determines whether or not two data points in a set of data belong to the same class based on their proximity to one another .

#### 3.4.2. Support Vector Machine:

Support vector machines are an effective form of an application of ML that is used for regression and classification analysis. Drawing a hyperplane based on the data points that most effectively separate the data points is how it works. The name given to this line is the "decision line." Those who are on one side of the decision line acknowledge the validity of the points made by those on the other side .

#### 3.4.3. Logistic Regression (LR):

It is one of the most widely used ML algorithms and is classified under the umbrella of "supervised learning approaches." A predetermined group of independent variables is used in conjunction with it to make a prediction about a categorical dependent variable. The outcome of a dependent variable that is categorical can be predicted using logistic regression. As a consequence of this, the findings ought to be either discrete or categorical. It's possible for the answer to be yes or no, zero or one, honest or dishonest, etc. But, rather than giving exact values between 0 and 1, it gives values that could fall anywhere between those two numbers.

#### 3.4.4. Random Forest:

The Random Forest machine learning algorithm is extremely versatile and user-friendly, and it pro-duces excellent results even without the need for hyperparam-eter tuning. This algorithm is constructed using the divide-and-conquer method, which results in the creation of a forest that contains multiple decision trees. Here, the root nodes and feature nodes are generated at random, and the maturity level of each tree is increased to its max-

imum potential based on the parameters that are already in place. When it comes to the training of predictive models, random forests are able to handle missing values very effectively.

#### 3.4.5. Gradient Boosted (GB):

Each predictor in GB works to improve on the one that came before it by lowering the amount of error it produced. On the other hand, the fascinating concept behind gradient boosting is not to fit a predictor to the data at each iteration but rather to fit a new predictor to the residuals produced by the previous predictor. This is done in place of fitting a predictor to the data at each iteration.

#### 3.4.6. LGBM:

LightGBM is a gradient boosting framework using tree-based learning algorithms. It is designed to be decentralized and efficient, with the following benefits: The training speed is fast and the efficiency is high,Low memory usage,Improves accuracy,Supports parallel and GPU training.It can handle huge amounts of data.

#### 3.4.7. CNN:

CNN is a deep learning algorithm that can take an input image, assign importance (learnable weights and biases) to different aspects/objects in the image, and distinguish them from each other. ConvNets require much less preprocessing than other classification algorithms. Filters are hand-designed in their own way, but with enough training, ConvNet can learn these filters/features.

## 4. Results and Discussion

### 4.1. Posterior Probability

The confusion matrix can be used to assess how successful a model is at resolving a classification issue. As a consequence of this, the output label is regarded as being binary because there are just two distinct possibilities for how it will turn out. This matrix is a grid that displays the distribution of the cases that were included in the test set that were predicted, and it does so according to class (TP, FP, TN, and FN in this case). Table I shows the components of the confusion matrix associated with this issue.

• True positive (TP): Many patients are at risk of stroke and are classified into the correct group by the classification model.

• True Negatives (Tn): The percentage of patients who are statistically very unlikely to suffer a stroke over the next few years, as determined by a classification model.

• False Positives (FP): The number of patients who were not at risk of sroke but were classified as at risk by the classification algorithm.

• False Negative (FN): The number of patients at risk of stroke who were classified as not at risk by the classification model.

Assess model correctness, precision, recall, and F1 score, using the different classifiers. The results obtained are shown in Table. II. Like a photograph Fig. 4,5,6,7,8,and 9 show the detection results graphically.

## 5. METHODS

**TABLE 2. Performance analysis of different classification**

| Model | Accuracy | Precisi | Recall | F1 Score |
|---|---|---|---|---|
| LR | 77% | 0.79 | 0.73 | 0.76 |
| RF | 93% | 0.92 | 0.31 | 0.93 |
| KNN | 93% | 0.93 | 0.92 | 0.93 |
| SVM | 85% | 0.91 | 0.78 | 0.84 |
| XG Boosting | 93% | 0.92 | 0.89 | 0.93 |
| LGBM | 92% | 0.91 | 0.92 | 0.91 |
| CNN | 98.9% | 0.97 | 0.96 | 0.98 |

### 5.1. Discussion

In this study, we show how different machine-learning algorithms can use a number of physiological factors to correctly predict a stroke. 1DCN outperforms all other algorithms with an accuracy of 98.9%. Table.III shows a comparison of the accuracy achieved by the different algorithms.

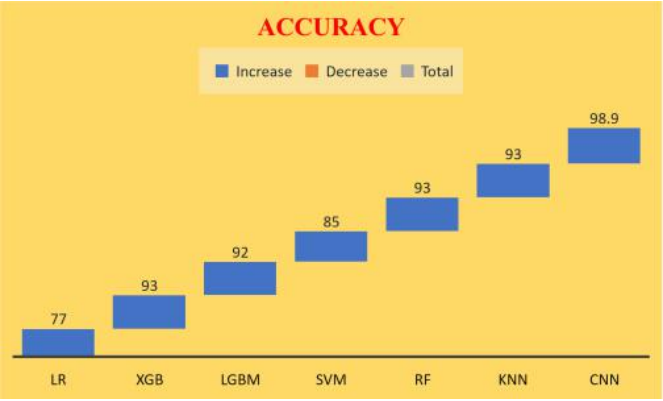
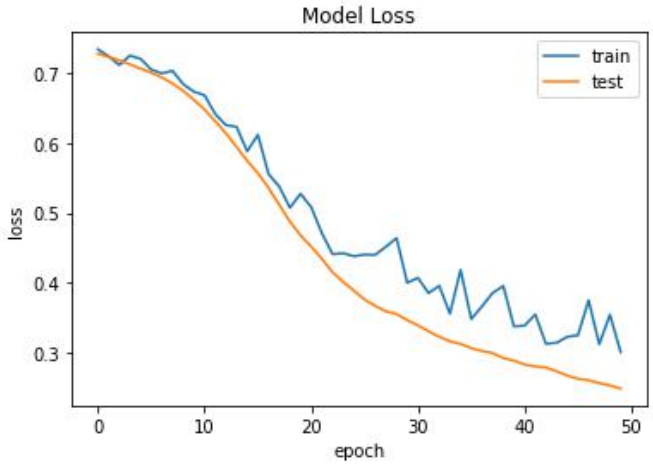
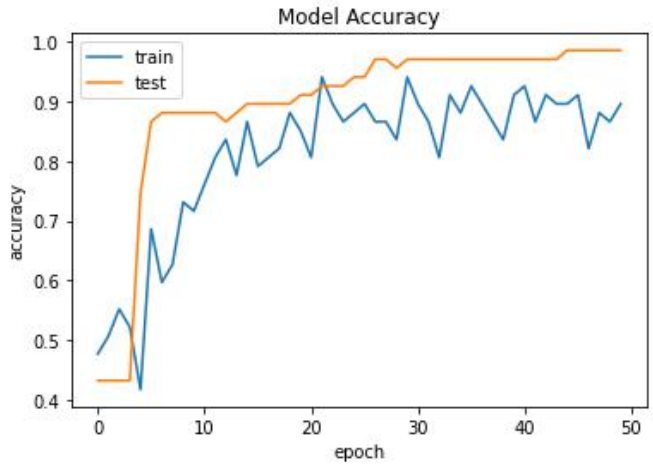



**FIGURE 4. Accuracy**




**Graph 1:** Graph of Model Loss




**Graph 2:** Graph of Model Accuracy

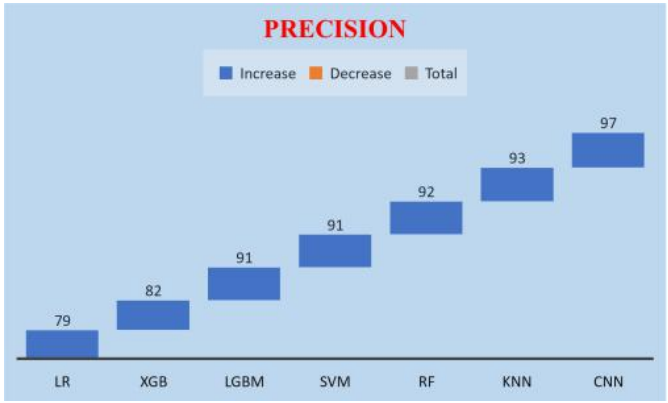



**FIGURE 5.** Precision

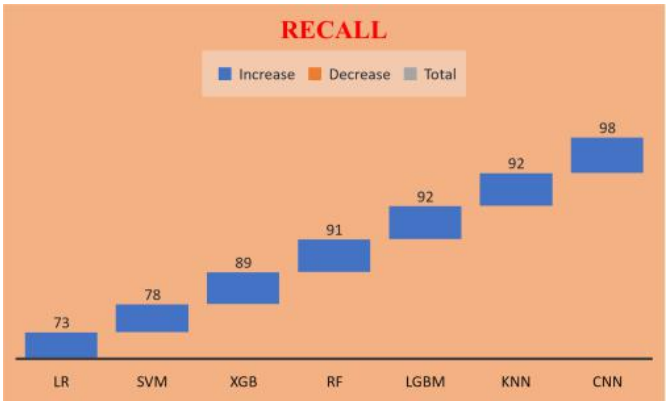



**FIGURE 6.** Recall

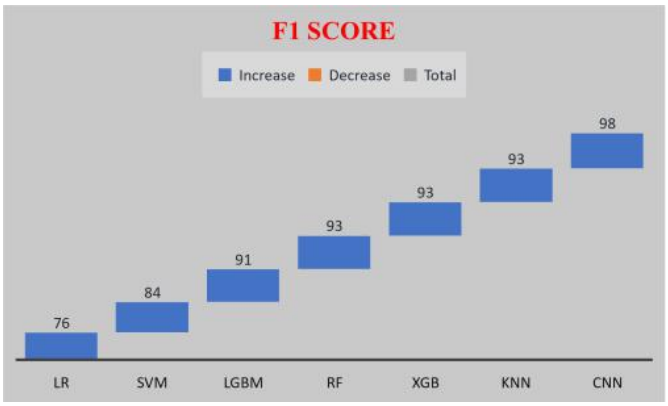



**FIGURE 7.** F1 Score

## 6. Conclusion

Heart stroke (HS) is one of the most life-threatening health risks, so many more people should find work in the medical field. This HS prediction model can be used to save lives and reduce stress in the workplace by reducing the amount of work that needs to be done by physicians and nurses, as well as by identifying patients who are at risk as a first step in the screening process. can boost output. To accomplish what we set out to do, we begin by applying a machine learning model to the well-known Framingham data set, and then we use a random forest classifier. The next step is to improve the classification accuracy by using cleaning, filtering, and feature selection and to monitor how this changes the results. The results of the experimental analysis are presented for various confusion metrics, including precision, F1 score, recall, and precision. After organising the data and ensuring that it was consistent across the board, we discovered that 1DCNN had the best performance. It has the highest degree of dependability.

## 7. The Goal of Future Research

According to the findings of a number of different machine learning techniques, it is clearly evident that the classification accuracy of 1DCNN models is insufficient for the practical application of this method. As a result, in the not-too-distant future,

we will be able to select which layers to use in deep learning models and which features to use in machine learning techniques by making use of a variety of different optimization techniques.

**TABLE 3. Comparison with the existing state-of-the-art Methods**

| Ref | Of Technique | Accuracy |
|---|---|---|
| (Fang, Huang, Z. Wang, et al.) | Convolutional Neural Network (CNN) | 83% |
| (Tougui, Jilbab, Mhamdi, et al.) | Artificial Neural Network (ANN) | 88.6% |
| (Emon et al.) | Weighted Voting Classifier (WVC) | 97% |
| Proposed | 1 Dementional CNN (1DCNN) | 98.9% |

## References


Akter, Bonna, et al. "A Machine Learning Approach to Detect the Brain Stroke Disease". *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (2022): 897–901.

Anitha, S, N Sridevi, et al. "Heart disease prediction using data mining techniques". *Journal of analysis and Computation* (2019).

Babič, František, et al. "Predictive and Descriptive Analysis for Heart Disease Diagnosis". *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems* (2017): 155–163.

Babu, Dheepitha, et al. "Gui based prediction of heart stroke using artificial intelligence". *Materials Today: Proceedings* 47 (2021): 104–108.

Beunza, Juan-Jose, et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)". *Journal of Biomedical Informatics* 97 (2019): 103257–103257.

Chaurasia, V, S Pal, et al. "Early prediction of heart diseases using data mining techniques". *Caribbean Journal of Science and Technology* 1 (2013): 208–217.

Emon, Minhaz Uddin, et al. "Performance Analysis of Machine Learning Approaches in Stroke Prediction". *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (2020): 1464–1469.

Fang, Gang, Zhennan Huang, Zhongrui Wang, et al. "Predicting Ischemic Stroke Outcome Using Deep Learning Approaches". *Frontiers in Genetics* 12 (2022): 2022–2022.

Fisher, Charles K, Aaron M Smith, Jonathan R Walsh, et al. "Machine learning for comprehensive forecasting of Alzheimer's Disease progression". *Scientific Reports* 9.1 (2019): 13622–13622.

Jalajajayalakshmi, V, V Geetha, M Mohammed Ijaz, et al. "Analysis and Prediction of Stroke using Machine Learning Algorithms". *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)* (2021): 1–5.

Mohan, Senthilkumar, Chandrasegar Thirumalai, Gautam Srivastava, et al. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques". *IEEE Access* 7 (2019): 81542–81554.

Patil, P B, P M Shastry, P Ashokumar, et al. "MACHINE LEARNING BASED ALGORITHM FOR RISK PREDICTION OF CARDIO VASCULAR DISEASE (CVD)". *Journal of critical reviews* 7.09 (2020): 836–844.

Sharma, Neeraj, et al. "Comprehensive Analysis of Feature Selection on Early Heart Strok Prediction". *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)* (2021): 142–147.

Sharma, Neeraj, et al. "Heart Stroke Risk Analysis: A Deep Learning Approach". *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)* (2021): 543–598.

Tougui, Ilias, Abdelilah Jilbab, Jamal El Mhamdi, et al. "Heart disease classification using data mining tools and machine learning techniques". *Health and Technology* 10.5 (2020): 1137–1144.

Verma, Luxmi, Sangeet Srivastava, P C Negi, et al. "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data". *Journal of Medical Systems* 40.7 (2016).

Wang, Meng, Xinghua Yao, Yixiang Chen, et al. "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients". *IEEE Access* 9 (2021): 25394–25404.



**Embargo period:** The article has no embargo period.

**To cite this Article:** , Mallikarjunamallu K, and Khasim Syed . "Stroke prediction using 1DCNN with ANOVA." International Research Journal on Advanced Science Hub 05.05S May (2023): 368–375. http://dx.doi.org/10.47392/irjash.2023.S050