International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Stroke Prediction Using Linear Regression

Nahala M A[1], Sooraj Subhash[2], Kishore Xavier[3] , Rahul Manoj[4] Sreehari V V[5]

[1]*Asst. Prof , Dept of CSE, Sree Narayana Gurukulam College Of Engineering, Kochi, India, nahalama@sngce.ac.in*
[2]*Student, Dept of CSE, Sree Narayana Gurukulam College Of Engineering, Kochi, India, soorajsubhash369@gmail.com*
[3] *Student, Dept of CSE, Sree Narayana Gurukulam College Of Engineering, Kochi, India, kishorexavier69@gmail.com*
[4] *Student, Dept of CSE, Sree Narayana Gurukulam College Of Engineering, Kochi, India, rahulmanoj2002@gmail.com*
[5] *Student, Dept of CSE, Sree Narayana Gurukulam College Of Engineering, Kochi, India, sreeharivinod666@gmail.com*

Abstract - Stroke is one of the leading causes of death and disability worldwide, and early prediction can significantly improve patient outcomes through timely interventions. This study explores the potential of using linear regression models to predict the likelihood of stroke in individuals based on a set of clinical and demographic factors. Data used came from a publicly available stroke dataset; the features used include age, gender, hypertension, heart disease, marital status, type of work, smoking habits, among others. The goal of this study is to find some important predictors and then establish a linear regression model which is capable of approximating stroke risk with reasonable accuracy.
Hence, feature selection and preprocessing aided the choice of relevant variables with which to build the predicting model. The subset formed by training and testing will be used to analyze a range of metrics, such as the mean squared error and the value of R-squared to reflect performance. The outcomes do indeed show that using relevant features for linear regression results can indeed be used for predictions related to stroke risks: thereby resulting in a simple but readable early risk identification model. More, however, the accuracy found of the model suggests that other algorithmic and data needs could allow for increased reliability in this field. The paper concludes that with linear regression, there seems a viable foundation to predict stroke while suggesting further refinement and more refined models would be necessary in clinical applications.

***Key Words*:** Stroke prediction, linear regression, feature selection, data preprocessing, machine learning.

1.INTRODUCTION

Stroke is one of the most leading causes of death and disability, and early prediction will be significantly improved through timely interventions. The aim of this study is to determine whether a linear regression model can be used for the prediction of stroke probability in a person given a set of clinical and demographic factors. Using a public available dataset concerning strokes, the variables present included age, gender, high blood pressure, heart problems, marital status, employment status, smoking habit, among others. It attempts to find the strong predictors in a linear regression model capable of giving a reasonable stroke prediction accuracy.

Stroke is a medical condition characterized by the sudden interruption of blood flow to the brain, resulting in loss of brain function. It is one of the leading causes of death and long-term disability worldwide, affecting millions of people annually. The ability to predict stroke risk is important for early intervention, prevention, and personalized treatment strategies that may reduce the burden of this debilitating disease. As healthcare systems shift towards making decisions based on data, predictive modeling has emerged as a promising tool for predicting medical conditions, including stroke.

Traditional stroke risk assessment is based on clinical guidelines and risk factors such as age, hypertension, diabetes, heart disease, smoking, and family history. These are all very well-known risk factors, but the interaction between them makes it challenging to quantify and predict stroke risk in individual patients with a reasonable degree of accuracy. Recent advances in machine learning and statistical modeling offer new opportunities for improving predictive accuracy. Linear regression is a very popular and interpretable statistical method that offers a straightforward approach to modeling the relationship between stroke risk and various demographic, medical, and behavioral factors.

This study aims to evaluate the utility of linear regression for stroke prediction. By using a dataset of individuals with various demographic and health characteristics, the goal is to develop a model that can estimate the likelihood of stroke based on these factors. Specifically, the study will focus on identifying which variables have the most significant impact on stroke risk and how well linear regression can predict stroke occurrence based on these inputs.

The primary impetus for this research is to investigate whether a simple, interpretable linear model can provide useful insights into stroke risk, which could become a potential tool for early screening in clinical settings. Although such complex models as logistic regression or machine learning techniques may have better accuracy, the study focuses on linear regression because of its simplicity, interpretability, and potential for real-time clinical applications. Ultimately, this research will contribute to the efforts in stroke prevention by showing the feasibility of predictive modeling in healthcare and forming a basis for further research with more complex algorithms..

2.STROKE PREDICTION

Stroke is a medical condition resulting from interruption of blood supply to the brain, leading to cell death. Accurate predictions of stroke can be vital in the early diagnosis and prevention of this disease, thus reducing mortality and disability rates. The prediction of strokes often depends on the analysis of risk factors such as age, hypertension, diabetes mellitus, smoking habits, and family medical history. Due to its capacity to detect patterns and inter-relations within a large dataset, machine learning techniques such as linear regression are extensively applied for this reason. Linear regression is a statistical model that relates the dependent variable with independent variables. In the context of stroke prediction, it computes the probability of having a stroke as a function of the input features, which represent the risk factors. The method is simple, computationally efficient, and interpretable, making it a good choice for understanding how individual factors influence stroke risk. The dataset typically contains records of patients' demographic information, medical history, lifestyle factors, and stroke outcomes. Preprocessing steps like handling missing values, normalizing data, and encoding categorical variables are crucial to ensure the dataset is ready for analysis. Feature selection is also very important for identifying the most relevant predictors and to remove noise in the model. The coefficients are computed by linear regression for every predictor variable, indicating how much the predictor contributes to stroke risk. Metrics like Mean Squared Error (MSE), R-squared, and residual plots are used to determine the accuracy and reliability of the model. While linear regression is useful for identifying risk factors, its simplicity may limit accuracy in predicting strokes, especially when relationships between variables are non-linear or complex. To address this, linear regression can be combined with the advanced methods of logistic regression, decision trees, or neural networks to improve predictive performance. Another promising area is the integration of real-time health monitoring data and personalization of medicine.

3.LITERATURE REVIEW

This paper “Stroke Prediction Using Machine Learning Classification Methods” by Hamza Al-Zubaidi, Mohammed Dweik, and Amjed Al-Mousa explores the use of machine learning techniques to predict stroke risk. Stroke is a major global health concern, and early identification of at-risk individuals is critical for effective prevention. The study examines classification models like Random Forest, Decision Tree, Logistic Regression, and Support Vector Machines (SVM), trained on features such as age, glucose levels, smoking habits, and medical history, to predict stroke occurrence. One of the challenges addressed in this study is the problem of imbalanced datasets, where stroke cases are significantly fewer than non-stroke cases. To address this, the authors use the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic examples for the minority class. This ensures the models are better at identifying stroke cases and reduces bias toward the majority class. Among the models tested, the Random Forest classifier showed the best performance, achieving an accuracy of 94-95% along with high precision, recall, and F1-score. The model’s ensemble approach, which combines multiple decision trees, contributes to its robustness and ability to outperform other methods like Decision Tree, Logistic Regression, and SVM. This study demonstrates that machine learning, particularly Random Forest, can serve as a powerful tool for stroke prediction. By integrating these predictive models into healthcare systems, medical professionals can identify high-risk individuals and take preventive measures effectively. This research also emphasizes the importance of addressing data imbalances to ensure reliable real-world applications.[1]

This paper "Early Stroke Prediction Using Machine Learning" by Chetan Sharma, Shamneesh Sharma, Mukesh Kumar, and Ankur Sodhi discusses the use of machine learning techniques for predicting stroke, with a focus on health and lifestyle factors. The classifiers involved in the study were Random Forest, Decision Tree, and Naïve Bayes. Random Forest showed the best accuracy of 98.94%, which indicated how good it was at classifying risky individuals. Feature selection from the

study indicated that features such as "gender" and "residence type" had minimal impacts on the predictions, but rather health-related factors such as glucose level and smoking habits. This is consistent with previous literature that focuses on the key medical variables for accurate predictions of strokes. The random forest has outperformed other methods because of the robust ensemble approach it uses; it processes more complex patterns better than other simpler models like decision tree and Naïve Bayes. This reliability makes it a promising tool for integrating machine learning into proactive healthcare systems. The study underlines the transformative role of machine learning, especially Random Forest, in advancing early stroke prediction. By adopting these models in healthcare, practitioners can enhance early detection and implement timely interventions, reducing the burden of stroke-related complications.[2]

This paper "Stroke Risk Prediction Model Using Machine Learning" by Nugroho Sinung Adi, Richas Farhany, Rafidah Ghina, and Herlina Napitupulu presents the use of the application of machine learning algorithms for predicting stroke risk. Because stroke is a leading global cause of death, the search is for an improvement on early detection by identifying persons with high risks through their historical health data. Machine learning is used to analyze the patterns in the data and enables the development of models for effective and accurate stroke prediction.The study evaluates three algorithms—Naïve Bayes, Decision Tree, and Random Forest—and compares their performance in predicting stroke risk. Random Forest was the most accurate model with a reliability of 94.78%, followed by Decision Tree at 91.91% and Naïve Bayes at 89.98%. With the ensemble method of Random Forest which takes several decision trees, it is better suited for datasets containing complex information. The results suggest that Random Forest demonstrates superior capabilities to handle large information and cannot be easily corrupted by noisy and incomplete inputs.The main strength of this study is that it provides a detailed comparison of different models, allowing insight into their strengths and weaknesses. Naïve Bayes and Decision Tree can do reasonably well, but they are outperformed by Random Forest in both accuracy and robustness. Thus, the study emphasizes choosing appropriate algorithms when developing machine learning-based diagnostic tools, particularly for high-stakes applications like stroke prediction.The authors also point out potential avenues for improvement, and include more patient attributes like genetic factors, lifestyle habits, and real-time health monitoring data may be included to enhance the performance of the Random Forest model. They recommend other advanced algorithms to test in order to check if even higher accuracy rates are possible in further studies.This research reinforces the potential of machine learning in healthcare, particularly in addressing life-threatening conditions like stroke. By integrating Random Forest models into medical systems, healthcare providers can achieve more reliable early detection of stroke risks, enabling timely interventions and significantly reducing stroke-related mortality and complications.[3]

The paper "Stroke Prediction Using Machine Learning" by Abinandhini D. M., Aman Kumar, Gudi Vishnu Teja, Divya S., Naman Chauhan, I. R. Oviya, and Kalpana Raja examines the stroke risk prediction of machine learning algorithms with the help of patient health data. The analysis was carried out on the following five models using a Kaggle dataset, with attributes including age, gender, BMI, and smoking status: Random Forest, Gaussian Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). From the study carried out, Random Forest has an accuracy of 94.81%, which outperformed all other models. In contrast, KNN was the model that delivered the lowest accuracy at 76.32%. Random Forest, because of its ensemble approach, handles large datasets and complex relationships well, making it the most reliable tool for the early detection of stroke. Other models, such as Gaussian Naïve Bayes and Logistic Regression, have done fairly well but are less robust compared to Random Forest.The analysis shows the strengths and limitations of each algorithm. While SVM and Logistic Regression have moderate accuracy, Random Forest has robust performance. KNN is lower in accuracy, which is an important point when the algorithms chosen for a particular dataset and prediction goal may not be suitable for medical applications. The authors suggest that machine learning models, such as Random Forest, should be included in healthcare systems to make early diagnosis and intervention for stroke possible. This method can improve patient outcomes considerably by facilitating early prevention measures. It further recommends further studies that consider other patient-related information such as real-time monitoring and genetic factors, which will fine-tune the models even better. This study validates the role of machine learning in healthcare, suggesting its capacity to transform the prediction of early strokes and furthering healthy outcomes.[4]

The paper “Stroke Prediction Using Machine Learning Methods” by Saumya Gupta and Supriya Raheja explores the use of machine learning algorithms to predict stroke risk, highlighting the importance of early detection for better outcomes. Using health and demographic data, the study evaluates algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Naïve Bayes,

and Random Forest to identify the most effective model for stroke prediction.Results show that Random Forest outperforms other models, delivering the highest accuracy due to its robust ensemble approach, which handles complex data and variability more effectively than simpler algorithms. Other methods, including SVM and KNN, show moderate accuracy but fall short in comparison to Random Forest's reliability.This study emphasizes the potential of machine learning in improving stroke diagnosis and enabling timely interventions. Integrating models like Random Forest into healthcare systems can aid in early detection, significantly reducing stroke-related complications.The authors encourage further research to enhance these models by including additional data such as real-time health monitoring and genetic information, which could improve predictive accuracy and make these tools even more effective in clinical settings.[5]

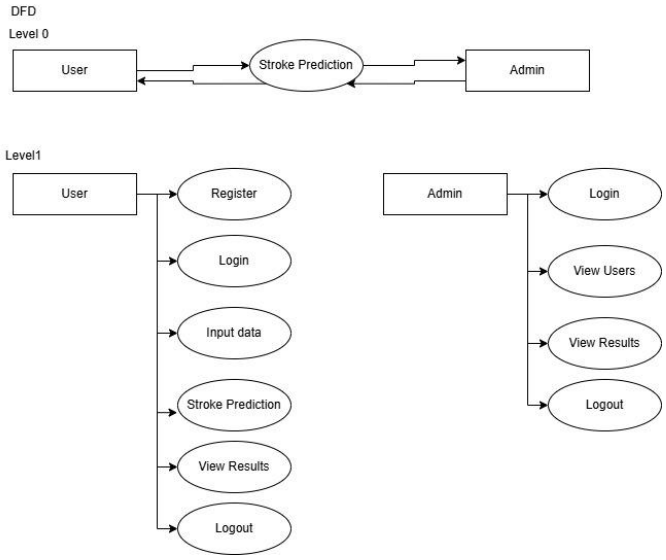
The paper "A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks" by Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, and Nishtha Jain offers a technique for stroke prediction by processing 29,072 patient electronic health records (EHR). The research highlights age, heart disease, hypertension, and average glucose levels as the most predictive features in the prediction of stroke.The authors use Principal Component Analysis (PCA) to reduce the dimensionality of data for efficiency in the model. They find that neural network on four features gives the highest accuracy. This shows the significance of proper feature selection in improving prediction.When compared to other machine learning models, the neural network consistently outperforms the others, showing superior results in stroke prediction. This underscores the value of deep learning for analyzing complex, large datasets like EHR.This paper points out the fact that this approach enhances clinical decision-making, because more accurate stroke risk assessments will allow for more timely intervention. This method should also simplify EHR management and further help healthcare professionals to appropriate use of patient data.The study emphasizes the scope of stroke predictability and early diagnosis through neural networks and machine learning from healthcare systems, which aids in further research in reinforcing these models.[6]

The paper "Prediction of Brain Stroke Using Machine Learning Algorithms and Deep Neural Network Techniques" by Senjuti Rahman, Mehedi Hasan, and Ajay Krishno Sarkar presents a machine learning approach to predict brain stroke using a Kaggle dataset. The study evaluates various machine learning algorithms such as Random Forest, XGBoost, AdaBoost, LightGBM, SVM, KNN, Naive Bayes, and Logistic Regression, along with deep neural networks (3-layer and 4-layer ANN) to assess their predictive capabilities.The highest accuracy is 99%, and it outperforms other algorithms in terms of classification accuracy, F1-score, and AUC. Here, it gets reflected that Random Forest is good at dealing with complex data and surpasses most other models.
The 4-layer ANN actually works pretty well within deep learning models and yields an accuracy of 92.39% while the 3-layer ANN shows to perform even worse than that. Although the deep learning model is promising, it does not outperform Random Forest in this application and therefore may rather be said that traditional ML methods could be better suitable for stroke prediction.The study focuses on the effectiveness of Random Forest in stroke prediction, noting its high accuracy and robustness. The paper concludes by recommending Random Forest as the most reliable model for clinical stroke prediction, while also recognizing the potential of deep neural networks for future research.In summary, this study compares a wide range of machine learning and deep learning methods, providing valuable insights into selecting the best predictive model for stroke detection, ultimately contributing to more accurate and timely clinical decision-making.[7]

A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks," authored by Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, and Deepu John, aims to predict the risk of stroke using an EHR dataset of 29,072 patients. The paper identifies age, heart disease, average glucose level, and hypertension as the most important predictive features for stroke prediction.To optimize the model, the authors apply Principal Component Analysis (PCA) for dimensionality reduction, showing that using only these four features leads to better accuracy than using all features. This shows how feature selection is important in improving model performance.The study compares several models, such as decision trees, random forests, and convolutional neural networks (CNNs). The neural network model outperformed the others, showing its effectiveness in stroke prediction because it can identify complex patterns in large datasets.The research highlights the importance of optimized feature selection in improving the accuracy of prediction. The neural network model significantly improved stroke risk prediction by focusing on the most relevant features.In conclusion, the study suggests that the combination of machine learning and neural networks could improve clinical decision-making in predicting stroke. Optimized models will help identify high-risk patients earlier, resulting in better outcomes. Further research should be conducted to refine the models and enhance their predictability.[8]

4.ARCHITECTURE

The Architecture of a Stroke prediction using linear rergression is:



The image represents a Data Flow Diagram (DFD) for a stroke prediction system. It illustrates how data flows between the system's entities and processes. The diagram is divided into Level 0 (a high-level overview) and Level 1 (a more detailed breakdown of system functionalities). Below is a detailed explanation

The Level 0 diagram, also referred to as the context diagram, is a high-level view of the stroke prediction system, with a focus on its interaction with external entities and the main process. It identifies the flow of data between the system and its key stakeholders. The Context Diagram or Level 0 comprises a comprehensive, general description of the stroke prediction system, focusing on its interaction and involvement between the central process of the system and the major elements involved. The two central persons involved in the stroke prediction system are the user and the admin. Each is meant to play a separate, yet interrelated, part in the functioning of the system. The user includes people who interact with the system to determine the possibilities of having a stroke. These users interact with the system by inputting personal information and health-related data; for example, medical history, age, or details about lifestyle, which then the system processes to predict. The system uses such data in predictive models that analyze it, and this usually comes back to the user in the form of outcomes, normally representing the probabilities of having a stroke. This process empowers users by providing them with valuable insights that can help them make decisions related to health or urge them to seek medical advice. In contrast, the admin plays the role of an administrator who oversees the system for its smooth operation and to maintain the integrity of the process. The admin logs in to the system to monitor user details and review stroke prediction results that are created for the individual. This oversight allows the admin to monitor system performance, address any anomalies, and potentially use the data for research or reporting purposes. At the core of the system is the stroke prediction process, which acts as the central mechanism facilitating these interactions. This process takes the input provided by the user, applies sophisticated algorithms or machine learning models to analyze the data, and generates accurate stroke risk predictions. The data flow in the system starts from users submitting their health data, which then gets processed by the central stroke prediction mechanism to yield results. At the same time, the admin can have access to those results and user information for management or analysis purposes. Therefore, the Level 0 diagram captures the basic structure of the stroke prediction system, highlighting the flow of data between users, administrators, and the prediction process. It is a basis for understanding at a general level how the system operates, hence of great utility to stakeholders and developers.

The Level 1 (Detailed Diagram) explains the procedures for both users and administrators in the stroke prediction system. On the user's side, it starts from the interaction using the register function to get an account that requires the entry of the user's personal credentials in order to log into the system. Then, it proceeds to login by introducing the credentials that will lead to the availability of the features offered by the system. After login, the users input their data, which is important health information such as age, blood pressure, and medical history. The next process is the stroke prediction process, where the system uses predictive models to analyze the data input by the user and determine the likelihood of a stroke. After the analysis, users can log in to view results. Users are given their specific prediction results, showing the probability of stroke occurrence for them. After checking the results, users can log out safely from the system. In the admin section, administrators are very important. Admin logs into the system by using the login function with the administrative details. Once logged in, they can access the view users feature, where they can view and manage all registered users' details. Moreover, admins can use the view results feature to see the stroke prediction results for all users, which is helpful for monitoring the system, research, or reports. Once done with the above administrative activities, the admin securely logs out of the system using the logout function. This detailed diagram will highlight how different user and administrator structured workflows may look, giving a

reflection of how this system properly manages data entry, process prediction, and the result managing in order to make experience smooth and operational for any stakeholder.

The stroke prediction system is designed to help users assess their risk of stroke based on health-related inputs, giving them valuable insights into their potential risk. It also offers administrators the ability to oversee and manage the system, including access to user data and prediction results. The structured flow of the system ensures smooth interactions for users and efficient administrative management. The Level 0 diagram represents a general view of the system, showing how it relates to external entities, such as users and administrators. The Level 1 diagram is more specific in nature, showing the functions and processes for both users and admins and demonstrating the logical sequence of actions within the system. This approach is very crucial in aiding developers and stakeholders understand the system's design and functionality in the planning and development stages.

5.COMPARISON WITH EXISTING SYSTEM

The current stroke prediction systems rely on a broad range of statistical models and machine learning algorithms that examine health information to give prognostication of stroke risk. These range from traditional methods, such as linear regression, logistic regression, which provided a foundation for stroke forecasting, to more complex or sophisticated machine learning techniques of decision trees, random forests, neural networks, SVMs, and KNNs. All these methods bring their individual strengths: classic models fit for more simple assignments, the advanced models better dealing with vast amounts of information and taking into account much more intricate and non-linear relationships of risk factors between each other. Linear and logistic regressions have widely been in use for many decades because they are so easy and straightforward for interpretation. They enable medical practitioners to understand the correlation between a patient's individual risk factors, such as age, blood pressure, and cholesterol levels, and their chance of having a stroke. These models are more practical when working with small data sets because they do not require large volumes of data to produce reasonable predictions. More advanced techniques, including decision trees and random forests, are able to manage larger and more complex data sets. These algorithms function by making a series of decision rules that split the data into various branches, aiding in the discovery of interactions between multiple risk factors. Though these models are highly efficient at identifying patterns, they are prone to overfitting if not tuned properly. Yet, another very advanced technique has developed neural networks, which would recognize rather complex patterns while processing voluminous data via multiple levels of interconnected nodes. Again, these are very much useful in high-scale health records or scanning, such as CT and MRI scan images but are computationally expensive to run and therefore require good hardware resources. Moreover, neural networks are often regarded as "black box" models because they are not very interpretable, and so it is hard for health care providers to understand why a prediction was made. SVM and KNN are also commonly used in the stroke prediction task. SVMs are particularly good for classification problems where the classes are well-separated, like distinguishing between patients at high risk for stroke and those not at risk. KNN is a non-parametric method that predicts outcomes based on the proximity of a given data point to its neighbors. While KNN can be intuitive, it struggles with large datasets, as the algorithm becomes computationally expensive when dealing with high-dimensional data. While these machine learning models and statistical techniques have proven useful in stroke prediction, they all come with certain disadvantages. These include issues related to the complexity of the models, the need for large datasets, high computational demands, challenges with interpretability, and difficulties with generalization to new populations.

Current stroke prediction systems rely on complex modeling with neural networks and random forests, while their output in most cases remains opaque and difficult for healthcare providers to understand and trust. This is a major issue in clinical practice: clear insights into risk factors are required. In contrast, the proposed system uses linear regression, which is highly interpretable and provides transparent insights into how factors like age and blood pressure impact stroke risk, hence making it even more suitable for clinical environments. In addition, advanced models like neural networks and SVMs are very computationally expensive and may not be feasible in the resource-poor healthcare environment, and their slow training phase and long implementation time further disqualifies their practical usage. The proposed system uses linear regression, which is computationally efficient and can run on less powerful hardware for real-time predictions in smaller or resource-constrained healthcare environments. Large volumes of data are required to train complex models effectively, however. Such datasets may not always be accessible, especially in smaller institutions. It addresses this proposition by using linear regression since it can work nicely with smaller datasets and ensures wider accessibility. Moreover, though such advanced models are likely to overfit, especially when there is unbalanced or inadequate data, the simple nature of linear regression deters the probability of overfitting if data is relatively well-balanced. Another challenge with advanced models

is that their high computational demands make it hard to scale; however, the proposed system can easily be retrained or recalibrated as more data becomes available and grows with the expanding dataset. In addition, complex models require expensive hardware for training and deployment, which is financially unfeasible for many healthcare providers. By employing linear regression, the costs outlined above are minimized and more economical, which is a solution that proves to be valuable for especially those areas where resources are limited. Existing models would, by then, lack generalizability to the new populations either based on demographic differences or regional variability, while linear regression, being less rigid, may easily get updated with newer data or risk factors in order to remain relevant for diverse clinical environments.

The proposed stroke prediction model has several advantages over a complex machine learning model due to its interpretability and efficiency, among other points. In contrast to complex models that are hard to track, the system uses the linear regression technique, whose explanation of how a risk factor, such as age or high blood pressure, affects prediction of stroke makes it sound more acceptable and reliable to patients and healthcare providers. In addition, the linear regression model is computationally efficient and performs well on less powerful hardware, which is critical in real-time clinical applications, where resources may be scarce, thus significantly better than the more resource-intensive models. The design is also simpler, and hence, there is a cost savings on the high-end machine learning infrastructure, and thus, it's a much more affordable solution for healthcare providers. This system will enable healthcare providers to flag high-risk patients early in order to target preventive interventions on such patients, and ultimately reduce the rate of strokes and improve patient outcomes. In addition, it will be easy to update the system, adding new risk factors or data sources with changes in medical knowledge; otherwise, more complex systems, once built, may require major reconstruction if new data are to be accommodated.

The proposed system depicts several distinct advantages over the ones in existence today, including simplicity, interpretability, computationally efficiency, and lower cost. In contrast to existing systems which depend heavily on advanced models of machine learning, for example neural networks, random forests or support vector machines, despite their power they pose serious challenges to themselves. These systems are highly complex, usually operating as "black box" models, and are thus hard to interpret in their decision-making processes. The decision-making process needs to be interpretable because of the clinical acceptance involved. Also, they need a lot of computational resources that are unavailable in most health settings, especially low-resource ones. In contrast, the proposed system uses a much more straightforward and interpretive technique, linear regression, where it is quite easy to understand how the different risk factors, for example, age, blood pressure, and lifestyle, affect the probability of stroke. Interpretability is also critical for gaining the trust of healthcare professionals and patients. Further, linear regression is computationally efficient and can be run on less powerful hardware, which makes it suitable for real-time clinical applications, even in resource-constrained environments. The system's simpler design also significantly reduces costs associated with high-end machine learning infrastructure, which makes it an affordable solution for healthcare providers with limited budgets. It is also scalable, capable of handling growing datasets and adapting to new risk factors as medical knowledge advances. Focusing on simplicity, interpretability, efficiency, and cost-effectiveness, the proposed system would be a practical, affordable, and scalable solution for stroke prediction, especially in healthcare facilities in resource-limited settings, overcoming most of the challenges posed by more complex existing systems.

6.CONCLUSION

In conclusion, the stroke prediction model developed with linear regression has a significant potential for improving stroke prevention and health care outcomes. It can easily identify individuals at risk using various health and lifestyle factors and is a very practical and interpretable solution for clinicians. The methodology followed in creating such a model begins with clearly defining the problem and identifying key factors influencing stroke risk. Data collection and preprocessing are two crucial stages in ensuring that the health data used for training the model is clean, relevant, and accurate. Once the data is processed, the model is built, trained, and evaluated to ensure that it accurately reflects the relationships between risk factors like age, blood pressure, cholesterol levels, and comorbidities such as diabetes or heart disease. This approach ensures that healthcare professionals can make informed decisions about which patients need further attention and preventive measures.

One of the major advantages of employing linear regression in stroke prediction is its interpretability. Linear regression does not produce such complex models as that of more advanced machine learning algorithms. Instead, it enables the clinicians to understand how a risk factor contributes to a particular prediction. This results in the enhancement of clinical acceptance and trust, given that the results can easily be explained to the patients. Success would therefore greatly rely on the quality and relevance of the data

used. Of course, inclusion of relevant predictors, such as lifestyle factors, age, and a more detailed medical history, may ensure the accuracy of model predictions. Further feature engineering and hyperparameter tuning improve the performance of the model in capturing relationships between the variables. The downside to this is that linear regression models are incapable of capturing complicated, nonlinear relationships between variables. So while linear regression works for many situations, the interaction of risk factors in complex ways might limit its use in other cases.

Beyond that, and to improve on the current model's accuracy in forecasting, other models may need to be introduced, especially logistic regression, which in some medical fields, have been replaced by machine learning techniques for instance decision trees, forests, or even support machines. Such techniques can reveal more complex relations and therefore possibly lead to higher predictability in cases of strokes. Despite this, the simplicity and interpretability of linear regression make it a great starting point for stroke prediction systems, especially in resource-limited settings. This model can give healthcare providers an effective, scalable, and real-time solution to determine who is at risk, allowing for early intervention. Ultimately, while linear regression can be a powerful predictor for stroke, the overall combination of such models with superior ones and continuous refinement against new data and research can go on to result in highly effective and personalized healthcare intervention with better outcomes for the patient in the long term.

7.REFERENCES

[1] Chetan Sharma , Shamneesh Sharma , Mukesh Kumar , Ankur Sodhi " Early Stroke Prediction Using Machine Learning " Chitkara University, Himachal Pradesh (INDIA).

[2] Chetan Sharma , Shamneesh Sharma , Mukesh Kumar , Ankur Sodhi " Early Stroke Prediction Using Machine Learning " Chitkara University, Himachal Pradesh (INDIA).

[3] Nugroho Sinung Adi , Richas Farhany , Rafidah Ghina , Herlina Napitupulu "Stroke Risk Prediction Model Using Machine Learning "

[4] Abinandhini D M , Aman Kumar , Gudi Vishnu Teja , Gudi Vishnu Teja , Divya S , Naman Chauhan , I R Oviya , Kalpana Raja " Stroke Prediction Using Machine Learning "

[5] Saumya Gupta , Supriya Raheja " Stroke Prediction Using Machine Learning Methods "

[6] Manikandan, K., Patidar, A., Walia, P., & Roy, A. B. "Hand Gesture Detection and Conversion to Speech and Text."

[7] Hwang, E. J., Cho, S., Lee, J., & Park, J. C. "An Efficient Sign Language Translation Using Spatial Configuration and Motion Dynamics with LLMs."

[8] Gong, J., Foo, L. G., He, Y., Rahmani, H., & Liu, J. "LLMs are Good Sign Language Translators."