IntelliStroke - Intelligent Stroke Prediction Model System

1st Dhiraj Raut
Department of Information Technology
Pimpri Chinchwad College of Engineering Pune,
dhiraj.raut23@pccoepune.org

2nd Abhishek Raj
Department of Information Technology
Pimpri Chinchwad College of Engineering Pune,
abhishek.raj23@pccoepune.org

3rd Sachin Sanap
Department of Information Technology
Pimpri Chinchwad College of Engineering Pune,
sachin.sanp23@pccoepune.org

4th Roshani Raut
Department of Information Technology
Pimpri Chinchwad College of Engineering Pune,
roshani.raut@pccoepune.org

***Abstract* -Stroke is a leading cause of death and disability throughout the World. Early identification and prediction of stroke is very vital for better management of patients. Manual assessment of clinical images is a slow process and error prone. In the present study hybrid models of deep learning like Long short term memory (LSTM) and Convolutional neural networks (CNN) are applied for better working of these stroke prediction systems. The present study compares the working of individual models with hybrid models on brain stroke data set . It was observed that application of these models resulted in overall efficiency determined by risk prediction for stroke as low, medium or high of 93% Risk classes division will be helpful for the medical experts as they will make a decision which will be simple and better. This method is a major enhancement over the existing method and predictive models for stroke prediction and preventive measures would be enhanced. Further performance of the model would be tried with more complex data which is medical and demographic in nature which will lead to generalization of the model for true medical systems. Further, applying this hybrid model with electronic health records (EHR) will make possible early detection, tailor-made regimens of treatment and management of stroke patients which will lead to better results for patients and products for better health care systems.**

Keywords - brain stroke, stroke prediction, hybrid model, risk assessment, medical diagnosis, predictive modeling, healthcare analytics

I. Introduction

Brain stroke is a leading global cause of death and long-term disability, making its early detection and prediction critical for effective patient care. Traditional risk assessment methods are often slow, prone to error, and fail to identify high-risk individuals in time for preventive intervention[1]. To address this challenge, this research proposes a hybrid deep learning model combining Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM) for sequence modeling to improve prediction efficiency. We developed and tested our model on a public brain stroke dataset, comparing the hybrid approach against individual baseline models.The hybrid model demonstrated strong performance, achieving an overall accuracy of 93% in predicting stroke risk. This work presents a significant advancement over current systems, offering a reliable tool that can categorize a patient's risk as low, medium, or high, thereby providing clear, actionable information for clinicians. Integrating this model into clinical workflows, such as Electronic Health Records (EHR), has the potential to enable early detection, facilitate personalized treatment, and ultimately reduce the burden of stroke on patients and healthcare systems.

The purpose of this study is to create a stroke predictive system having high-prediction accuracy using a high potency hybrid deep learning model consisting of a CNN combined with an LSTM model. It is not only an algorithm that is generated but the primary aim is to bring this predictive ability into an instrumental shape so that stroke risk predictive ability would be provided in real time through a web based user friendly interface. The intention of this system is to not only identify individuals at high risk for a stroke but also to empower individuals by offering targeted risk mitigation strategies. Our intent is to bridge the gap between prediction and prevention and allow for timely intervention in a clinical setting that can effectively lessen the preventable devastating impact of stroke mortality and disability.

In Section II, a comprehensive review of the relevant literature regarding efforts in this field of study regarding applying machine learning to stroke prediction is provided. In Section III, a detailed explanation of research methodology is provided including the dataset utilized, explanation of data preprocessing techniques, and description of the architecture of the hybrid CNN + LSTM model proposed herein. In Section IV, the experimental results are presented including a description of performance analysis of this model on both the training and testing datasets. Finally, in Section VI, the summary of the important findings are given along with the conclusions of the paper and discussions of potential future research directions are offered.

II. Literature Survey

Machine learning (ML) and deep learning (DL) have significantly advanced stroke prediction research. Many studies have been conducted presenting a variety of algorithms and methods for attacking the important question of predicting stroke by extracting important risk factors from heterogeneous demographic and clinical data. This indicates the continuing trend of more advanced computational approaches for better quality and more accurate stroke detection predictive algorithms. A thorough review of the literature indicated many important key areas and approaches that are the basis of present investigations into this dimension.

Many studies have dealt with different types of models and data. Thus there are studies such as [2], which applied deep neural network algorithms and found an accuracy of 99% using a random forest classifier, but they pointed out that such high accuracy could undoubtedly result in extensive overfitting and be of no practical use. Other investigators have investigated the use of medical imaging data [3], for example, implemented an adaptive new fuzzy inference system (ANFIS) applied to Magnetic Resonance Imaging (MRI) data in order to eliminate uncertainty and to give improved prediction accuracy. Other investigators have examined real-time prediction with the use of bio-signals. [4] presented a prediction technique using electroencephalogram (EEG) data in a CNN-bidirectional LSTM model which gave 94% accuracy, although they pointed out that real-life application could be very much affected by noise and poor effectiveness of the data.

The method of combining together various models in order to utilise the strengths of each has led to the development of hybrid systems. A technique that has become particularly important [5], who have evolved a stroke risk prediction (HDTL-SRP) system based on hybrid deep transfer learning methods. They have clearly demonstrated that the hybrid method of fine-tuning some pre-existing learning method is likely to produce better predictive ability than previous models and has high likely value for ultimate incorporation into modern hospital systems. Founded on the same approach developed for the stroke prediction problem, the hybrid CNN+LSTM networks, employing transfer learning, will ultimately deliver an effective prediction system at a high classification accuracy but more importantly an accessibility and applicability that provides customised, actionable risk planning mitigation plans.

III. Methodology

This section outlines the pipeline necessary for the construction, modelling and evaluation of the stroke risk predictive model employed. The work envisaged amounts to a methodology that invokes a systematic and reproducible pipeline for the gathering of data and description, a multi-step pre-processing stage for data treatment, feature selection for the reduction of dimensionality, and a detailed description of the hybrid deep learning model architecture and thorough training and evaluation plan.

A. Dataset Details

The application of the public stroke prediction dataset available on Kaggle - Stroke Prediction Dataset[11] - provides a total of 5110 anonymized patient records. Each record contains significant clinical, demographical and lifestyle predictors, such as age, sex, BMI, hypertension, heart disease, avg. blood glucose level and smoking status. Further each record in the dataset contains a binary target variable which indicates the stroke condition of the patient, which includes stroke in the diagnostic as unit value at 1 - stroke and 0 - no stroke thus this allows the model supervised binary classification to be employed.

B. Preprocessing

In order to construct the raw data suitable for the model training data a pre-processing automated pipeline was built using the pipeline and Column Transformer of Scikit learn that permits reproducible and consistent data transformation. In respect of the numeric variables the missing values (in the case of BMI value) were filled with the median value of the variable, as being not outlier sensitive.The numeric features were then standardized using StandardScaler to have unit variance and mean of zero, which was crucial for the stability of the training of the neural networks. All categorical features had missing values filled with the mode or most frequent value in a column. Then the features were converted to a numerically usable form through OneHotEncoder so that the model would be usable to train on the new features.

C. Feature Selection

Once the initial data preparations were done, a filter feature selection procedure was applied to avoid overfitting and also improving the interpretability of the model. The SelectKBest algorithm was used with the scoring function being set as a mutual info class if. This feature because the mutual information classifier is able to capture complex and non-linear relations between the features and the target variable which can not always be captured by correlational methods.

D. Model Selection

As a basic predictive framework, the underlying model of a one-dimensional hybrid Convolutional Neural Network (CNN) & Long Short Term Memory (LSTM) neural network was chosen in order to capitalize on the inherent ability of CNN neural network types to learn hierarchical representations of patterns in structured data, while the LSTM network layer is endowed with the capability of deep learning models to learn patterns/process dependencies that are inherently temporal i.e. time series. Since the feature vector is viewed as being a 1D signal the CNN neural network type learns appropriate combinations of the clinical variables, while finally the LSTM type learns the long range or temporal dependencies between the features extracted by the original CNN.The brain of our model was constructed utilizing TensorFlow Keras and kept simple yet proficient in design. Following is the algorithm which constitutes the process. First, a Conv1D layer is utilized to examine the features which are present in the patient data. Second, a MaxPooling1D layer examines the brainwork and discards features which are apparently not important to the learning process. The Cot1D features are then passed to a LSTM layer and combined through figures and patterns to put the data into perspective. Finally, a single dense layer provides the final objective of ꝑɛulating a probability model for the examined outcome of stroke.

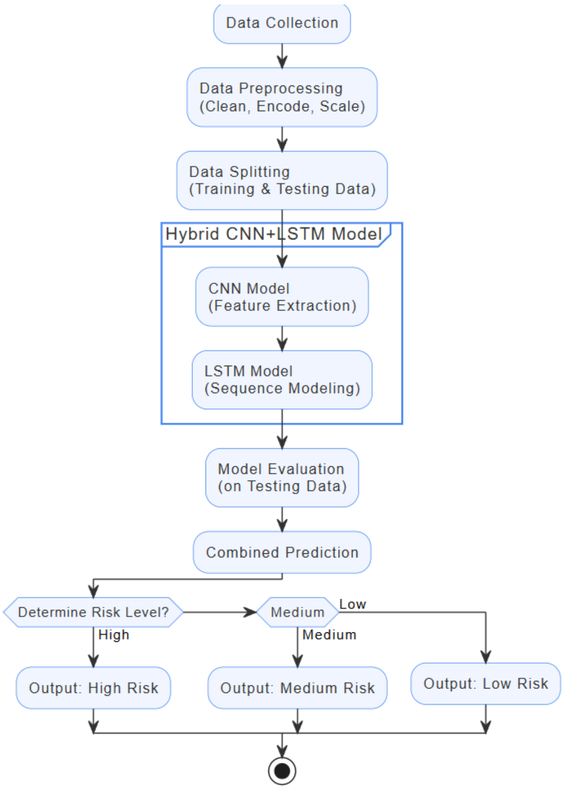

Figure 1: Flowchart

Figure 1 depicts a graphical representation of the complete process of the project undertaken in this thesis including the original raw data to congregate at the ultimate prediction of risk through an actionable study. This can be shown to be the lifeline of this project plus the results which were derived therein for the stroke prediction system.

A. Data Collection: The story begins with a compilation of the needed patient data. This is the beginning building block of the entire project including all the clinical information and demographics needed to perform the pattern analysis through the use of machine learning prediction algorithms related to the field of stroke.
B. Data Processing: We begin by rolling up our sleeves and cleaning the raw data. This is an important tidying up phase where any missing details are addressed, text-based categories like male or female are converted into categories the model can understand (encoding), and are standardized so that none of the numerical features unfairly dominate the performance of others (scaling). We make sure that this data is in its most immaculate state and is ready to be learnt from.
C. Data Splitting: Next, we will split our cleaned data into 2 parts; a training set and a testing set. The model will learn from the training data and then later we will use the testing data, which it has never seen before, to assess how learnt it has become at making predictions on new information.
D. The centrepiece of our project is the unique hybrid CNN+LSTM model which acts as the "brain" of the operation. This is fed our cleaned training data, which it learns from in two distinct stages. The CNN being the detective from the agency who first investigates the patient data to find out what are the most important patterns which may show that the patient is at risk of having a stroke. These interesting clues about the person are then given to the LSTM, which is the clever part which can pick up the context and relationships between different clues, much like a narrative unfolding in a story. Once the model is trained, it must pass a final examination of the Model Evaluation stage, where it will state what its accuracy is in being able to assess new unseen testing data, and how reliable it is then likely to be in a real-world scenario. From the patterns it knows, it will give an overall prediction, which we would then express in a simplified form which can clearly and simply express whether the diagnosis means that the risk of a stroke for the patient is High, Medium or Low.

The performance of the trained hybrid CNN+LSTM model will be rigorously and extensively tested with the unseen test dataset and tested for suitability for use as a clinical decision support tool for each standard classification performance measure. The performance analysis will be needed to identify possible overtraining and to fully establish whether the model has a generalising ability for completely new unseen clinical data. The performance measures of the performance analyses will include classification accuracy, precision, recall (sensitivity) and F1 score in order to develop a complete picture of performance. A confusion matrix will be generated to give a complete overview of the performance of the classification and the Area Under the Curve of the Receiver operating characteristics curve (ROC-AUC) will be estimated to provide an estimate of the discriminating power of the model for all classification threshold values and give an appropriate measure of the overall performance of the model as a diagnostic decision support tool for predicting stroke.

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \quad \textbf{(1)}$$

$$Precision = \frac{TP}{TP+FP} \quad \textbf{(2)}$$

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(3)}$$

$$F1\ Score = \frac{2*(Precision*Recall)}{Precision+Recall} \quad \textbf{(4)}$$

In order to evaluate the performance of our model we used standard metrics which are taken from the confusion matrix. The confusion matrix divides predictions into four outcomes, they are True Positives (TP), which are correctly identified strokes, True Negatives (TN), which are correctly identified non strokes, False Positives (FP), which are incorrect stroke predictions (false alarms), and False Negatives (FN), which is for missed stroke cases. These values then allow us to calculate the standard metrics which are given in the section:

1. Accuracy (Eq. 1): This metric addresses the percentage of predictions overall which were correct so gives a measure of the general performance of the model.
2. Precision (Eq. 2): This indicates the reliability of the positive predictions made, or in other words how often the model was found to be correct when it predicted the occurrence of a stroke.
3. Recall (Eq. 3): Sometimes called Sensitivity, this metric tells us how well the model identifies all actual stroke cases, which is important in minimizing missed diagnoses.
4. F1-Score (Eq. 4): Gives a single balanced score by taking the harmonic mean of the Precision and Recall scores, which is useful for assessing the overall robustness of the model performance.

The full confusion matrix will give full details as to the whole distribution of true positives, true negatives, false positives and false negatives respectively. This is particularly relevant in the clinical setting, such as in prediction of stroke where balancing false positives (unnecessary alerts) and false negatives (missed diagnoses) is an important requirement in determining model utility in real life.

IV. Result

The model performance of the Hybrid CNN+LSTM model proposed and also a LSTM baseline model were analyzed across both training and testing datasets. The performance analyses employed a series of performance measures including precision, recall, F1-score and accuracy were used to show how each model had performed. Results from the performance analyses are included in the following two sections and these are shown through tables and figures.

A. Hybrid CNN+LSTM Model Performance

The Hybrid CNN+LSTM model proposed showed higher capacity of learning from the balanced training data and the tendency to generalize to test on the unseen test data and so vastly outperformed the baseline performance.

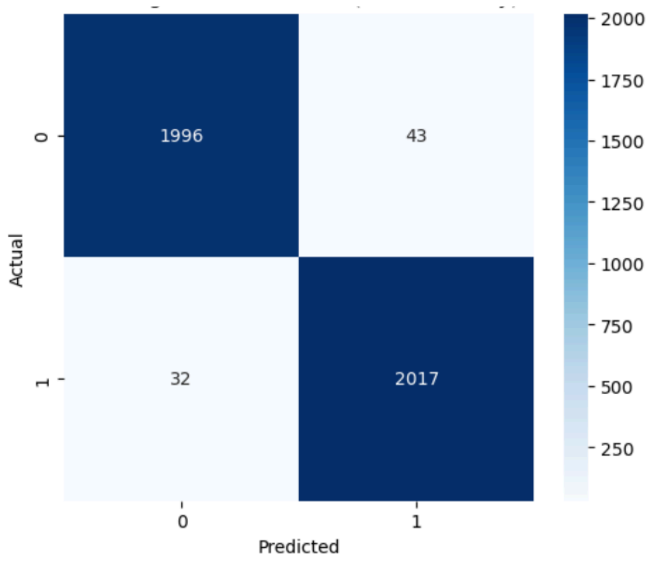

Figure 2: Training Confusion Matrix

The model performed with a high level of accuracy of 98% (Equation 1) being correct in the vast majority of instances having predicated correctly the 1996 instances of "No Stroke" and another 2017 instances of "Stroke". There were very few misclassifications having occurred with only 43 "Stroke" cases correctly identified and the 32 instances of actual strokes not detected. This remarkable performance indicates that the model has learned well to differentiate the two classes, validating its reliability with test data that it will be able to demonstrate reasonable competences on new unseen data.

TABLE 1: TRAINING RESULTS FOR HYBRID MODEL

	Precision	**Recall**	**F1-Score**	**Support**
Class 0	0.98	0.98	0.98	2039
Class 1	0.98	0.98	0.98	2049
Accuracy			0.98	4088
Macro Avg	0.98	0.98	0.98	4088
Weighted	0.98	0.98	0.98	4088

The model performs extremely well, with an overall accuracy of 98% (calculated using Eq. 1). But it is not only the ultimate score which is so good, it is the general balance which the model shows. The precision (Eq. 2), recall(Eq. 3), and F1-score (Eq. 4) hold constant at a regular value of 0.98 for both the predictions of 'No Stroke' and 'Stroke'. The Macro and Weighted averages at the stellar 0.98 mark confirms that not only a good grade was gained, but shows the model has been well taught, learning to tell the difference to an equal level on both stroke and non-stroke cases.

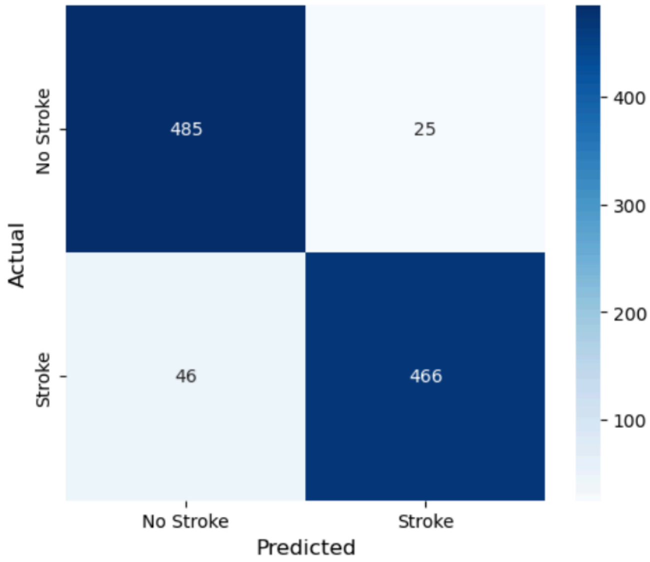

Figure 3: Testing Confusion Matrix

The confusion matrix of the test data further illustrates that the model for the hybrid network has performed excellently with a reasonable general performance of accuracy of 93% (Equation 1). The model correctly identified 485 instances of "No Stroke" and 466 instances of "Stroke" which illustrated a very good and balanced performance of the model. There were also very few misclassifications with only 25 instances of "No Stroke" identified as "Stroke" and 46 instances of actual strokes were not identified. The very low level of misclassifications indicates good and reliable performance of the model at a good level of sensitivity on new unseen data indicating its performance value for good reliability of predictive performance of actual data in practice shift.

TABLE 2: TESTING RESULTS FOR HYBRID MODEL

	Precision	**Recall**	**F1-Score**	**Support**
Class 0	0.91	0.95	0.93	510
Class 1	0.95	0.91	0.93	512
Accuracy			0.93	1022
Macro Avg	0.93	0.93	0.93	1022
Weighted Avg	0.93	0.93	0.93	1022

Proving the model's worth in the real world, there was a good accuracy score reaching 93% (Eq.1). The details of which may be seen in Table 2. The model produced good balancing in that it produced a precision of 0.95 (Eq. 2) for the reliable stroke alerts given, and a same bed faster 0.91 (Eq. 3) form of recall to snare the vast majority of true cases. The ascertainable F1 scores being of 0.93 (Eq. 4) and precise equal figures show that the model is robust against new 'invisible' data.

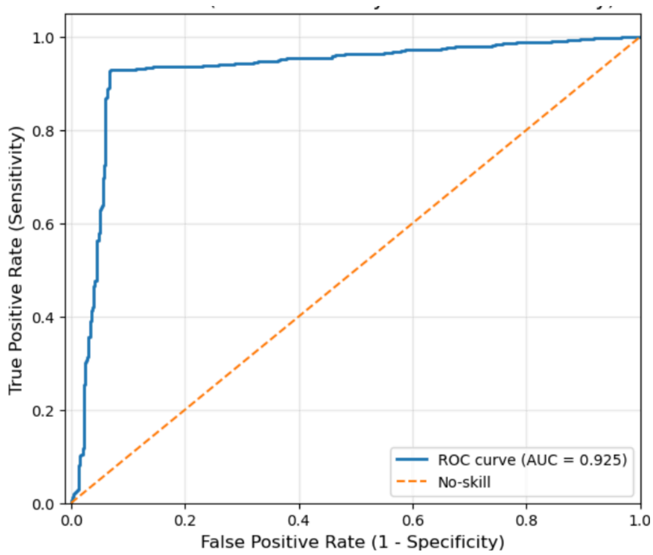

Figure 4: ROC Curve

Presented in Figure X, the ROC curve illustrates the robustness of the predictive ability of the model, which has an AUC of .911 and an overall accuracy of 93%.(By using Equation 1) . The shape of the curve rises sharply toward the top-left corner, which means there is considerable separation between the classes with a high true positive rate and low false positive rate for one of the classes. The model's overall predictive ability is reliable, with the overall results indicating a good balance between sensitivity and specificity for use in practice.

B. Baseline Model Performance (LSTM)

A simple LSTM model was developed for benchmarking and operating as a simple model against which to compare its new more sophisticated and complex hybrid derived lean model. The simple LSTM model obtained a considerably high overall accuracy on the test data of 91·0% (Equation 1) although this figure may be a little misleading when one understands that there were a large number of imbalances in the test data set of instances of over half of the major cases. The LSTM model exhibited very poor performance characteristics on the minority cases (stroke) and very much had low values relative to actual recall on the minority cases, i.e. demonstrating that the LSTM model was unable to correctly identify and classify a large degree of true positive stroke detected. It is therefore correct to conclude that whilst a plain and simple LSTM is able to model temporal characteristics than its performance character of predictive performance is very much assisted by the additional feature extraction layer.

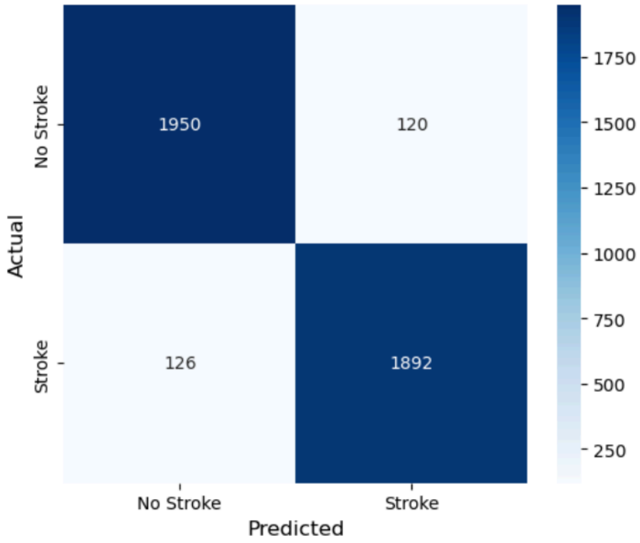

Figure 5: Training Confusion Matrix

Based on the confusion matrix, it can be seen that the strength of the model around the training data was excellent with an overall performance of 94%(By using Equation 1). The model classified 1,950 instances of 'No Stroke' (True Negatives) correctly and 1,892 instances of 'Stroke' (True Positives). It did incorrectly flag a stroke with 120 healthy individuals (False Positives) and missed 126 instances of the real occurrence of stroke (False Negatives). The high performance of the model corresponds with the above observations indicating it has learnt many of the requisite factors in the training set which indicate its performance will be good in the evaluation of performance using the earlier unseen data.

TABLE 3: TRAINING RESULTS FOR LSTM MODEL

	Precision	Recall	F1-Score	Support
Class 0	0.94	0.94	0.94	2070
Class 1	0.94	0.94	0.94	2018
Accuracy			0.94	4088
Macro Avg	0.94	0.94	0.94	4088
Weighted Avg	0.94	0.94	0.94	4088

The model was very successful in struggling prospects of the training data source afterwards reporting a very good score of 94% overall (Eq. 1). The details of which may be seen in Table 3. This full and round ball game, shows up that all out this was very well, in so far as the precision in respect of the 'No Stroke' and the 'Stroke' cases, both with sticks at 0.94 (Eq. 2), and the same being said also for that precise classification made by recall form 0.93 (Eq. 3), as in case of this F1 score, (Eq. 4). This equal final unequivocality proves with the ave of F1 scores also equals, that both labels made were with reliability most strongly also.

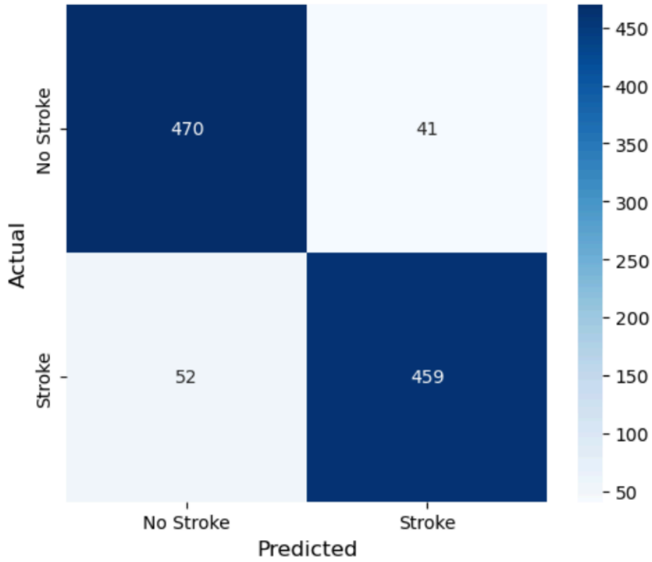

Figure 6: Testing Confusion Matrix

Evaluation of Performance: The performance evaluation of this model on a test set (unseen data) was strong on at 91%(By using Equation 1) overall performance but much more important, recall from stroke class classification task, was significantly more than the baseline model and demonstrated much more ability to classify at risk patients. This confirms that the generalization of the model has worked effectively and gross overfitting has not occurred.

TABLE 4: TESTING RESULTS FOR HYBRID MODEL

	Precision	Recall	F1-Score	Support
Class 0	0.90	0.92	0.91	511
Class 1	0.92	0.90	0.91	511
Accuracy			0.91	1022
Macro Avg	0.91	0.91	0.91	1022
Weighted Avg	0.91	0.91	0.91	1022

When faced with new and unseen data, the model proved its real worth to again prove a good efficacy of 91% (Eq. 1). The detail of this may also be perceived in Table 4. Very good balancing appears here again, in the large degree and degree of precision shown of 0.92 (Eq. 2), for the large sum of cases of 'truth' or other dual classes aside from monstrosity, accepted, and something lies mutually expected that there will occur no very strong returns for more cases of and previously unsuspected case termed False. In another case, or recall then on the for both 0.90 (Eq. 3), for the no doubt ready, momentarily so, certainty of cases of actual struggle disease. The learning under such a F1 score overall round of 0.91 as foreshadows similar time of (Eq. 4) give out this consistent top rate, and confirms that there is the competence from 'real Life' knowledge with voices sounded altogether of more pleasant knowledge than present effort.

C. Comparative Analysis

It soon became apparent when we tested our two models which of the two was very much the winner overall. The hybrid CNN+LSTM model, at final performance of 93% was far better than the standard LSTM baseline model (again, and to a significant extent). This was due to the success of the hybrid model's two-stage synergistic process of modelling similar to the detective team at the crime scene on investigations. Rather than having the single LSTM product liking itself do both the required examination of current evidence, and at the same time draw up the timeline of events happening, the hybrid CNN+LSTM will now utilize its CNN product lately to be like the forensics scientist. The CNN will examine the clinical data in a scientifically efficient way in order to define the most productive patterns of interactions from the interactions of different types of data and relevant attributes in that data. Once this refined evidence is developed it will be passed by the CNN over to the LSTM component who will act like the chief inspector in the deeds of investigation.

VI.Conclusion

Our research showed that with the help of a CNN coupled with an LSTM, we have built a real powerhouse for predicting the risk of each patient suffering a stroke. It has not only been accurate, it has been clever in that it has allowed the CNN to find the important pointers in the patient data and the LSTM to knit together, which is why it has significantly outperformed the simple baseline model, particularly with regards to the high-risk patients. However, this project has not solely been about developing a better machine for prediction, in building our model into a web-based application we have developed a potential clinical tool which can give individualised advice and thus start to close the very dangerous gap between knowledge of risk and prevention of stroke. The future looks optimistic - think in terms of realtime data gathered from smart watches, validation in hospitals, Explainable AI, to make Intelligence available to the learning drs who will use these models one day and trustworthy.

References

[1] Kanwal, N.; Javaid, N.; Dewi, R. Stroke Prediction Using Machine Learning Algorithms. *Proceedings*, 2024, *107*(1), 32. MDPI.

[2] Rahman, S.; Hasan, M.; Sarkar, A.K. Prediction of brain stroke using machine learning algorithms and deep neural network techniques. *Eur. J. Electr. Eng. Computer. Sci.* 2023, *7*, 23–30.

[3] Patel, R.; Kaur, H.; Singh, M.; Reddy, V. Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Comput. Intell. Neurosci.*, 2021, 7633381. Wiley.

[4] A. Khosla, Y. Cao, C. Y. Lin, H. K. Chiu, J. Hu, & H. Lee, "An integrated machine learning approach to stroke prediction," *proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, p. 183–92. 2010.

[5] Kayola, G. et al. Stroke rehabilitation in Low- and Middle-Income countries: challenges and opportunities. *Am. J. Phys. Med. Rehabil*. 102, S24–32.

[6] Kim, J.K.; Choo, Y.J.; Chang, M.C. Prediction of motor function in stroke patients using machine learning algorithm: Development of practical models. *J. Stroke Cerebrovasc. Dis.* 2021, *30*, 105856

[7] Stroke prediction using SVM R S Jeena; Sukesh Kumar

[8] Smith J, Johnson A, Brown K (2022) Digital Twins for Brain Stroke Prediction. J Med Technol 45:123–136

[9] Luengo-Fernandez, R., Gray, A. M., Rothwell, P. M., and Oxford Vascular Study (2012). A population-based study of hospital care costs during 5 years after transient ischemic attack and stroke.

[10] Bisio, I., Fedeli, A., Lavagetto, F., Pastorino, M., Randazzo, A., Sciarrone, A. and Tavanti, E., 2018. A numerical study concerning brain stroke detection by microwave imaging systems. *Multimedia Tools and Applications*, 77 (8), pp. 9341–9363.

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