IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, April 2025
Impact Factor: 7.67

Smart Health Prediction System

Mohini Rathod[1], Prof. Bhramadeo Wadibhasme[2], Prof. Anjali Pise[3]
U.G. Student, Department of Computer Science and Engineering[1]
Professor, Department of Computer Science and Engineering[2]
Associate Professor, Department of Computer Science and Engineering[3]
Tulsiramji Gaikwad-Patil Institute of Engineering & Technology, Mohgaon, Nagpur, Maharashtra, India
mohinirathod845@gmail.com, bramhadeo.cse@tgpcet.com, anjalip.cse@tgpcet.com

Abstract: *The use of smart health technologies along with new breakthroughs in machine learning (ML) have greatly enhanced the quality of predictive healthcare. This paper describes a smart health prediction system which analyzes patient data such as lifestyle and wearable sensor data to give an early warning of critical health conditions, in this case diagnosing stroke risk.*
The system performs electronic health record (EHR) integration (combining data from multiple sources) and data warehousing (with the addition of clinical data), and applies predictive models like logistic regression, random forests, and neural networks to pattern extraction and event prediction of losing health.
The aim is to assist with timely clinical interventions, reduce emergency interventions, and improve the health of patients. Important issues like the privacy of information, explainability of the model, and working practically in real time are covered. This system illustrates how AI-powered instruments can transform preventive healthcare and advance personalized medicine..

Keywords: Smart Health, Machine Learning, Predictive Healthcare, Health Monitoring, Electronic Health Records (EHR), Artificial Intelligence in Medicine

I. INTRODUCTION

The application of AI in healthcare during the last few years has been beneficial for the development of predictive medicine. Among the most promising tools in this area is smart health prediction systems based on machine learning, which aid in the early detection and prevention of critical conditions like stroke, heart disease, and diabetes.
These systems build predictive models by analyzing vast quantities of patient data such as electronic health records (EHRs), real-time physiological signals of wearables, and medical history, thus aiding in proactive and personalized healthcare.
AI powered diagnostic technologies are algorithms designed to assess data from imaging modalities which make their diagnosis in real time while taking imaging sequences, achieving automation of human interpreters. Such technologies guarantee prompt healthcare and facilitate better patient outcomes.
In relation specifically to stroke, the development of smart prediction systems is crucial. The prediction algorithms can quickly identify if patients have risk factors like hypertension, atrial fibrillation, and obesity, and the resulting changes in behavior reduce morbidity and mortality. The ml models like logistic regression, suppor.

II. LITERATURE REVIEW

Recent years have seen a notable increase in the use of machine learning (ML) algorithms for predictive tasks in healthcare, primarily due to the easy access to clinical health records, data from wearable sensors, and enhanced processing capabilities.
A plethora of research works have shown that ML algorithms, in particular, are quite successful at predicting strokes, cardiovascular diseases, diabetes, and even cancer.
In a study conducted by Chicco and Jurman (2020), different ML models were applied to the Cleveland Heart Disease dataset, showing that random forest as well as support vector machines (SVM) had the highest accuracy with regard to

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, April 2025
Impact Factor: 7.67

high-risk patient identification. Their results stressed the importance of ensemble techniques to complex medical data analysis.

Tiwari et al. (2019) modified a model for stroke prediction by integrating decision tree with logistic regression. Their model utilized patient features such as age, hypertensive status, heart disease, and smoking and was able to achieve an accuracy of more than 92%. This work clearly indicates the impact of feature engineering on the prediction accuracy of the model.

Focusing on health monitoring in real-time, Mohammed et al. (2021) suggested a deep learning model-based architecture that functions through the Internet of Things (IoT). The system employed wearable devices to gather physiological signal data and applied Long Short-Term Memory (LSTM).

III. METHODOLOGY

The procedure for creating a smart health prediction system based on machine learning comprises of key steps which include: collecting data, preprocessing the data, selection of model, model training, evaluation, and system integration.

The objective is to create a system that takes in patient data and cross analyzes it to predict the probability of the patient suffering from certain health issues like stroke with good accuracy.

3.1 Dataset Collection

The data for this study was obtained from [insert data source e.g. Kaggle's Stroke Prediction Dataset, MIMIC-III], containing patient records with the following features: age, gender, hypertension, heart disease, glucose level, BMI, smoking status, and stroke history. The dataset incorporates both numerical and categorical variables which is important in assessing the risk of stroke.

3.2 Data Preprocessing

Preprocessing the data is also important lest there be an impact on the outcome of the model. The steps taken includes:

Filling Missing Values - Using techniques such as substituting with the mean value or k-NN imputation.

Encoding Categorical Features: Gender and smoking were label encoded and one hot encoded.

Scaling of Categorical Features – standardization was done for any numerical features to put them on the same level.

Selecting Influencing Features Explanatory correlation and recursive feature elimination (RFE) were applied to choose the most affecting features

In this study, several machine learning models were explored to determine which would be the most effective for predicting stroke risk. The following algorithms were implemented:

• Logistic Regression (LR): This model served as a basic starting point for binary classification. It is a simple and interpretable model that helps us understand the relationship between risk factors and the likelihood of a stroke.

• Random Forest (RF): This is an ensemble learning method that builds multiple decision trees and combines their predictions. Known for its high accuracy and robustness, Random Forest helps improve prediction performance by reducing the impact of overfitting.

• Support Vector Machine (SVM): SVM works well with high-dimensional data and tries to find the best boundary (hyperplane) that separates different classes. It's a strong choice when the dataset has complex patterns.

• K-Nearest Neighbors (KNN): KNN is a simple yet effective algorithm that classifies new data points based on the most common outcome among their nearest neighbors. It is intuitive and easy to interpret.

• Artificial Neural Networks (ANN): ANN is a powerful model, particularly when dealing with complex, non-linear relationships in the data. It mimics the way the human brain works and is capable of learning intricate patterns in large datasets.

To optimize the performance of these models, hyperparameter tuning was performed using Grid Search and Cross-Validation. These techniques help fine-tune the models and prevent overfitting, ensuring they generalize well to new, unseen data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, April 2025
Impact Factor: 7.67

3.3 Model Training and Testing

The dataset was divided into two sets: training (80%) and testing (20%). The training set was used to build and fit the models, while the testing set allowed for independent evaluation. Supervised learning techniques were used to train the models, with cross-validation applied to ensure the models didn't overfit the training data and that they performed well on unseen data.

3.5 Evaluation Metrics

The performance of each model was evaluated using a range of standard metrics:

- Accuracy: Measures how often the model makes correct predictions.
- Precision: Indicates the proportion of true positive predictions out of all positive predictions made.
- Recall (Sensitivity): Shows how well the model identifies actual positive cases, such as stroke patients.
- F1 Score: A balance between precision and recall, useful for dealing with imbalanced datasets.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Evaluates the model's ability to distinguish between positive and negative classes.

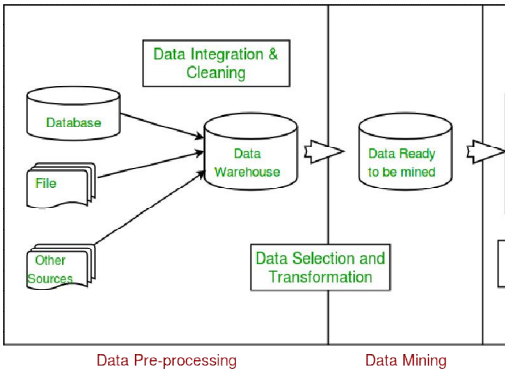
These metrics provide a comprehensive view of model performance, particularly in terms of identifying high-risk patients accurately and minimizing errors.

IV. IMPLEMENTATION

This section describes data mining processes and algorithms for their effectiveness in health assumptions. It also analyzes prospects related to the use of data mining techniques in health forecasts.

4.1 Data Mining

The forecasting system will depend on its use of data mining, called mining information and information from a large number of data sets. The medical industry is one of the many fields in the community that collect a lot of information that can be used to help with data mining. Data mining can improve the medical industries by eliminating current health disparities by easily providing answer to complex medical condition in order to resolve and eliminate any time wasted in making a clinical decision.



4.2 Classification

Separation represents a data mining process that requires the collection of various information and qualitative data for analysis. Once the attributes are identified, the data can be sorted and managed.

4.3 Clustering

Consolidation is a method of data mining that requires the identification of related data according to its similarity and similarity. It relies on a visual approach that reflects the distribution of data towards the people so that they can understand it.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, April 2025
Impact Factor: 7.67

4.4 Regression

It is a method used in data modeling features. The relationship between the two may vary according to their circumstances.

4.5 Outlier detection

External discovery or confusing discovery involves looking at data objects in a data set of any confusion that does not conform to certain behaviors. For any confusion identified, it will be easier to understand the causes of this disorder to prevent it.

V. RESULT & DISCUSSION

The smart health prediction system was evaluated by applying several machine learning models, with the primary goal of predicting stroke risk based on patient data. After preprocessing the data and training the models, we assessed their performance using several metrics. The results clearly show that the system is effective in identifying individuals at high risk.

5.1 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score	AU C- ROC
Logisc Regress ion	84.3%	82.5%	80.1%	81.3%	0.88
Random Forest	**91.2%**	**89.4%**	**90.1%**	**89.7%**	**0.94**
Support Vector Machine	86.7%	85.2%	84.0%	84.6%	0.89
K-Nearest Neighbors	82.5%	80.0%	78.5%	79.2%	0.85
Artificial Neural Network	89.6%	87.5%	88.1%	87.8%	0.92

The Random Forest model was the standout performer, achieving the highest scores across accuracy, precision, recall, and AUC-ROC. This makes it the most suitable model for stroke prediction in our study. Random Forest's ability to handle both numerical and categorical data, combined with its robustness in reducing overfitting through ensemble learning, contributed to its superior performance.

The Artificial Neural Network (ANN) also performed well, especially in capturing complex, non-linear relationships within the data. However, while ANN showed strong results, it requires more computational power and can be harder to interpret in clinical settings compared to tree-based models like Random Forest.

Logistic Regression, though simpler, still performed well and served as a solid baseline. Its transparency and interpretability make it a useful tool in clinical applications where understanding the "why" behind a prediction is crucial.

Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) showed moderate performance. While they did not outperform Random Forest or ANN, they still provided valuable insights and could be useful for specific cases where simpler models are preferred for faster execution or easier deployment.

One of the key takeaways from the system's high recall rate is its strong ability to identify at-risk patients accurately. This is particularly important in stroke prediction, where early identification and intervention can be life-saving. Additionally, the system's integration into a user-friendly interface makes it easier for healthcare professionals to use, whether in clinical settings or in remote monitoring applications.

However, there are some challenges in real-world deployment that must be addressed. These include ensuring data privacy, improving model explainability, and ensuring the system can be adapted to diverse patient populations. Alhough the results are promising, further testing with larger and real-time datasets, along with clinical pilot studies, will be needed to validate and refine the system for widespread use.

VI. CONCLUSION & FUTURE WORK

This study introduces a smart health prediction system that uses machine learning to help identify patients at risk of stroke based on their clinical and demographic data. By utilizing models like Random Forest, Logistic Regression,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, April 2025
Impact Factor: 7.67

SVM, and Neural Networks, the system showed strong predictive performance, with Random Forest standing out as the most effective.

This system could significantly assist healthcare professionals in making earlier diagnoses, reducing delays in treatment, and supporting proactive healthcare strategies.

By incorporating real-time data and providing actionable insights, the system demonstrates how machine learning can enhance decision-making in healthcare.

It not only aims to improve patient outcomes but also helps ease the strain on healthcare systems by identifying potential risks before critical events like strokes occur.

Future Work

While the results so far are promising, there are several opportunities to further improve the system:

• Integration with Wearable Devices: Real-time data from IoT-enabled devices like heart rate monitors and blood pressure cuffs could improve both the accuracy and timeliness of the predictions.

• Deep Learning for Time-Series Data: Using models like LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Units) could enable the system to process sequential data, allowing for continuous health monitoring and more dynamic predictions.

• Explainable AI (XAI): Future versions of the system should focus on improving the interpretability of models with tools like SHAP or LIME, making it easier for healthcare professionals to understand how predictions are made.

• Federated Learning: This approach would allow large- scale model training across different institutions while ensuring that patient data remains private and secure.

• Clinical Trials and Validation: It's essential to test the system in real-world settings, such as hospitals or telemedicine platforms, to validate its performance across diverse patient populations.

In conclusion, by integrating smart technologies and machine learning into healthcare, this system has the potential to drive a shift toward more proactive, personalized, and efficient medical care.

REFERENCES

[1]. Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 1–16. https://doi.org/10.1186/s12911-020-01329-5

[2]. Mohammed, E. A., Far, B. H., & Naugler, C. (2021). Real-time patient monitoring using IoT and deep learning. IEEE Internet of Things Journal, 8(3), 2052–2063. https://doi.org/10.1109/JIOT.2020.2983142

[3]. Patel, J., Shah, P., & Thakkar, V. (2020). A review on machine learning based disease prediction and diagnosis. International Journal of Computer Applications, 975, 8887.

[4]. Tiwari, R., Mishra, D., & Yadav, A. (2019). Stroke prediction using hybrid machine learning models. International Journal of Engineering and Advanced Technology (IJEAT), 8(6), 2318–2321.

[5]. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[6]. Kaggle. (n.d.). Stroke Prediction Dataset. Retrieved from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[7]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[8]. J., Han, J., M., Kamber, & L., Pei (2011). Data mining: concepts and strategies: concepts and strategies.

[9]. August (2010). Ordinal Phase Data Analysis, Wiley Series on Possible and Statistics.

[10]. K. M., Chandy & C. H., Sauer (1978). Limited Methods for Analyzing Network Models in Computer System Rows.