Machine Learning in Hospital Health Prognostication: Next-Gen Health Monitoring

Mrs. Poornima H.S[1], Madhalasa S.J[2] , Anushree Y.N[3] , Nakul Carpenter[4] , Shantanu Yadav[5] ,

[1]Asst. Prof. Dept. of AI&ML, RNSIT , VTU

[2,3,4,5] Students, Dept. of AI&ML, RNSIT , VTU

Abstract:

Accurately predicting the course of illness and potential outcomes for hospitalized patients remains a top priority in healthcare. Introducing a machine learning-driven health prognostication system for hospitalized patients. Leveraging diverse patient data, the model outperforms traditional methods in predicting adverse health events. Evaluation on a comprehensive dataset demonstrates enhanced accuracy and interpretability, offering valuable insights for clinicians. In this paper, methodologies proposed for combination of k- Nearest Neighbour algorithm, Convolution Neural Network and Neural Network featuring Recurrent Structure where both the momentary memory and vanishing gradient challenge and Gated Recurrent unit (GRU) processing is preferred to achieve accuracy of 92% and Area under Curve-Receiver Operating Characteristics (AUC-ROC) of 0.94 based on various merits like age, cholesterol.. This system has possibility to change the way patients are treated by fostering proactive interventions and refining resource allocation strategies. The inclusion of age, cholesterol, and other pertinent factors in the predictive model enhances its robustness and clinical applicability. This study introduces a machine learning-driven health prognostication system for hospitalized patients. Leveraging diverse patient data, the model outperforms traditional methods in predicting adverse health events. Evaluation on a comprehensive dataset demonstrates enhanced accuracy and interpretability, offering valuable insights for clinicians.

Keywords: Machine Learning; Advance algorithms; Data Analytics; Medical History; Healthcare.

1. INTRODUCTION

Fueled by the desire to enhance patient outcomes and optimize resource allocation, healthcare practitioners are increasingly turning to ML-driven approaches for prognostication [1]. Ensemble learning, a potent ML method that blends many models for enhanced prediction accuracy and robustness, has garnered significant interest in the last few years [2]. Studies have underscored the effectiveness of ensemble models in disease prediction across various conditions [3]. Further investigations have delved into the complexities of applying ensemble learning to predict multiple diseases simultaneously, emphasizing facets like model selection and performance evaluation [4].

Moreover, research has documented the efficacy of recurrent neural networks (RNNs) in detecting early signs of heart disease [5]. These studies collectively highlight the transformative potential of ensemble learning and ML in disease prediction and early detection. By integrating knowledge from various places, researchers aim to shape future research endeavors and help to contribute to the advancement of predictive healthcare analytics [6].

Applying machine learning for foreseeing disease underscores the escalating significance of advanced predictive analytics in contemporary healthcare [7]. Ensemble learning techniques, which synthesize the predictive prowess of multiple models, emerge as formidable tools in enhancing prediction accuracy across a spectrum of medical conditions [8]. By amalgamating insights from diverse sources, researchers showcase the potential for ensemble methodologies to refine disease prediction models, yielding more dependable prognostications [9]. Furthermore, the implementation of collaborative learning expedites early disease detection, notably in critical conditions such as heart disease, where timely intervention is pivotal for augmenting patient outcomes [10]. By dissecting subtle patterns in medical data, ML algorithms, such as RNNs, discern early disease indicators, enabling proactive measures to preempt risks and halt disease progression [11]. In addition to augmenting predictive accuracy and facilitating early detection, predictive analytics in healthcare offers significant dividends in optimizing resource allocation [12]. By pinpointing high-risk patients and projecting disease trajectories, healthcare providers can allocate resources judiciously, ensuring interventions are targeted where most imperative [13]. This not only elevates patient

care but also mitigates strain on healthcare systems and curtails expenses [14]. Moreover, the synthesis of findings from existing research informs future directions, propelling innovation in healthcare delivery [15]. The existing research identifies gaps in several areas: a lack of explicit gap analysis within the studied research, a need for improved feature selection methods to enhance model performance,uncertainties in multi disorders diagnosis using machine learning techniques, insufficient exploration of hybrid models for heart disease prediction, and a call for comprehensive data availability and adoption of advanced analytical methods for diseases like dengue and cervical cancer.

The accuracy can be improved on factors like precision, recall, AUC-ROC. This paper is organized with method, proposed method here algorithm combination is explained followed by result and discussions.

2. METHODOLOGY

The methodology includes gathering medical information gleaned from hospital files, processing it to handle missing values and normalize features. Then machine learning models such as KNN, CNN and RNN are trained using the data that has already been handled to predict patient health outcomes, aiding in clinical decision – making and prognostication.

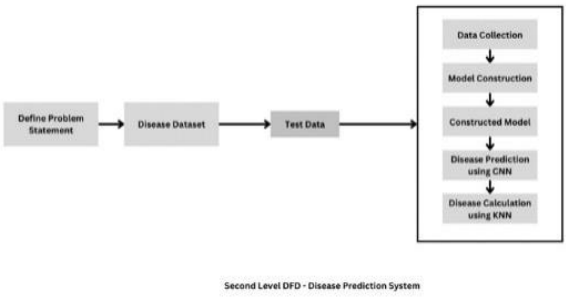


Figure 2.1: Block Diagram of Health Prognostication.

The block diagram in Figure 2.1 shows a second-level Health prognostication system. Here's a breakdown of the process in 3 lines:

1.The system starts with defining the problem statement and collecting data on the disease.

2.Then, it builds a model to predict the disease using a convolutional neural network (CNN).

3.Finally, it refines the prediction using a k-nearest neighbors (KNN) algorithm for disease calculation.

KNN: Suitable for smaller datasets and interpretable predictions. It finds the In the training collection k nearest neighbours and uses them to foresee the class label (classification) or value (regression) for an updated set of data. K-Nearest Neighbors (KNN) is a straightforward yet useful method utilised in supervised learning regarding challenges involving regression and classification. Unlike many other algorithms, KNN is non-parametric, i.e., it doesn't assume anything about the fundamental data distribution in the equation 1,

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d}(x_j - X_{i_j})^2]} \quad (1)$$

In our research, the Euclidean distance metric is primarily employed with the purpose of putting into practice the K-Nearest Neighbors (KNN) algorithm, facilitating the finding the data points that are closest.

The equations and terms that are commonly related to K-Nearest Neighbors are:

1.Manhattan Distance: Calculates The gap between two points by summing the absolute differences of their coordinates as in equation (2),

$$\boldsymbol{distance} = \sum_{i=0}^{n-1} |(\boldsymbol{x}[i] - \boldsymbol{y}[i])| \quad (2)$$

2. Minkowski Distance: A generalization of Euclidean and Manhattan distances, adjustable with a parameter r. The parameter is r, The quantity of dimensions is n (attributes) and p_k and q_k are respectively the kth attributes (components) or data objects p and q as in equation (3),

$$\boldsymbol{dist} = (\sum_{k=1}^{n} |\boldsymbol{p_k} - \boldsymbol{q_k}|^r)^{\frac{1}{r}} \quad (3)$$

3.Weighted Vote: Determines the impact of a point determined by the inverse square of its distance, giving closer points more weight as in equation (4),

$$vote_{weighted} = \frac{1}{d(p,q)^2} \quad (4)$$

4.Classifier Decision: Predict the class with the majority vote (or weighted vote) among the k-nearest neighbors.
5.Regression Decision: Predicts a value according to the values' average of its k-nearest neighbors as in equation (5)

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (5)$$

6.Choosing K: The process of Finding how many neighbours to take into consideration, often optimized through experimentation.
7.Normalization/Standardization: Adjusts the scale of data features, allowing for better comparison and convergence as in equation (6),

$$x_{new} = \frac{x-\mu}{\sigma} \quad (6)$$

8.Hamming Distance: Used for categorical data, counts the number of positions at which the corresponding symbols are different as in equation (7),

$$d(x,y) = \frac{1}{n}\sum_{n=1}^{n=n} |x_i - y_i| \quad (7)$$

9.Cosine Similarity: Calculate the angle of two vectors' cosine, used in high-dimensional spaces as in equation (8),

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (8)$$

CNN: Effective for learning complex patterns in image or time-series data, often used in medical imaging tasks. CNNs use convolutional layers to extract features, pooling layers for dimensionality reduction, and fully connected layers to discover the connections between the prognosis and the characteristics target as in equation (9),

$$E(v,h) = -\sum_{K=1}^{K} h^k \cdot (\varpi^k * v) - \sum_{K=1}^{K} h^k b_k - \sum_{i,j} h_{i,j}^k - c\sum_{i,j} v_{i,j} \quad (9)$$

Large datasets including tagged examples are used to train CNNs. The network modifies its internal parameters (weights and biases) by a procedure known as backpropagation to lessen the discrepancy between its predictions and the actual labels.

The expressions and equations which are frequently connected to Convolutional Neural Networks (CNN) illustrating foundational concepts and operations within these models are:

1.Convolution Operation: Applies a filter (kernel) over an input to produce a feature map, highlighting the specific features as in equation (10),

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (10)$$

2. ReLU Activation Function: Non-linear function that establishes all values as negative in the input to zero, promoting sparsity and efficiency as in equation (11),

$$f(x) = \max(0, x) \quad (11)$$

3.Pooling Operation (Max Pooling): Reduces the dimensionality of input by taking the maximum value over a window for each region of the feature map as in equation (12),

$$P_{ij} = \max_{l\in[i,i+k-1], m\in[j,j+k-1]} (X_{lm}) \quad (12)$$

4.Softmax Function: Converts the output scores from the network into probabilities by taking the exponentials and normalizing them as in equation (13),

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (13)$$

5.Cross-Entropy Loss: A loss function, commonly used in classification, that calculates The variance of a pair of probability distributions as in equation (14),

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \quad (14)$$

6.Filter (Kernel) Application: Detects patterns, such as edges or corners, by applying small matrices over the input data as in equation (15),

$$W_{ij}(I) = \frac{1}{|\omega|^2} \sum_{k:(i,j)\in\omega} \left(1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \epsilon}\right)$$
$$q_i = \sum_j W_{ij}(I) p_j \quad (15)$$

7.Feature Map: The result of applying a filter to the input, representing detected features at various locations as in equation (16),

$$F_{i,j} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{i+m,j+n} \cdot K_{m,n} + b \quad (16)$$

8.Strided Convolution: Reduces the output size by skipping input values by a stride length during convolution as in equation (17),

$$S(i,j) = (I * K)(si, sj) \quad (17)$$

9.Padding: Adds zeros or another padding surrounding the perimeter of the input, allowing filters to apply at the borders as in equation (18),

$$O = \frac{W-F+2P}{S} + 1 \quad (18)$$

10.Batch Normalization: Stabilizes learning by normalizing the inputs of each layer to have zero mean and unit variance as in equation (19),

$$\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} \quad (19)$$

3. PROPOSED METHODOLOGIES

This model diagram below in Figure 3.2 shows the process of collecting and analysing patient data. Here's a breakdown of the process:

1. Patient data is collected from three sources lab findings, computerised health records, and medical imaging.
2.The data is then analysed.

3.Finally, the analysed data is integrated and visualized on a dashboard.

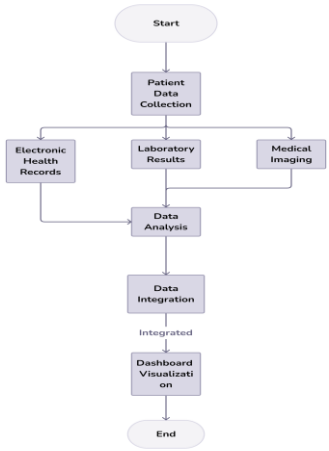


Figure.3.1 Model Diagram of Health Prognostication

The network diagram as shown in Figure 3.2 depicts a software development process. Here's a breakdown of the process:

- First, a database is created and data is collected. Then, a model is selected and generated.
- Next, the model is evaluated. If successful, the model is deployed. If not, the process loops back to testing different models (KNN, CNN, RNN) until a successful model is generated.

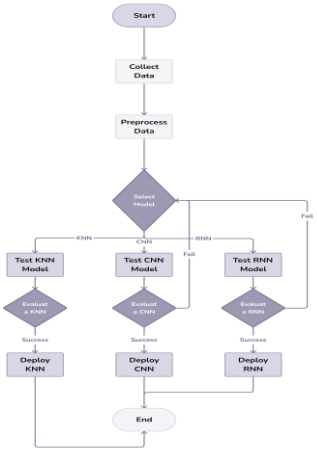


Figure 3.2 Network Diagram of Health Prognostication

Hear the block diagram as shown in Figure 3.3 shows a process for data acquisition and processing for disease prediction. Here's a breakdown of the process:

1. Data Acquisition: A patient's data is gathered by the system.
1. Data Processing: the data is examined to identify functional issues and symptoms.

2. Disease Prediction: Based on the information retrieved, a model is utilised to forecast potential illnesses.
3. Data Output: The most likely illness is generated by the system.

The text within the block labelled "Train_Data()" suggests this system might involve machine learning. A sizable collection of labelled examples is used to train machine learning algorithms. In this instance, patient data and related illnesses could be a part of the training set. The model gains the capacity to recognise patterns that distinguish between various illnesses throughout training.

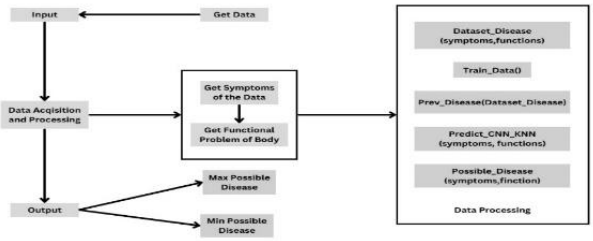


Figure 3.3 Architectural Overview

Mainly algorithm used for clasification of disease are :-

1.k-Nearest Neighbors (k-NN)

2.Convolutional Neural Network(CNN)

3.Recurrent Neural Networks (RNN)

Recurrent Neural Network (RNN): It is a category of artificial neural systems tailored for sequential data analysis. They are perfect for applications such as natural language, audio recognition, and time series prediction because of their design, which allows them can handle sequences of any length. RNNs are effective tools for modelling sequences because of their capacity to grasp temporal relationships in data with intricate patterns and dynamics as in equation (20),

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$y_t = W_{hy}h_t \qquad (20)$$

The scatter plot as shown in Figure 3.4 below, with "thalach" on the x-axis and "chol" on the y-axis, visually represents the relationship between maximal heart rate achieved during exercise ("thalach") and serum cholesterol levels ("chol"), shedding light on their potential influence on the likelihood of experiencing a heart attack.

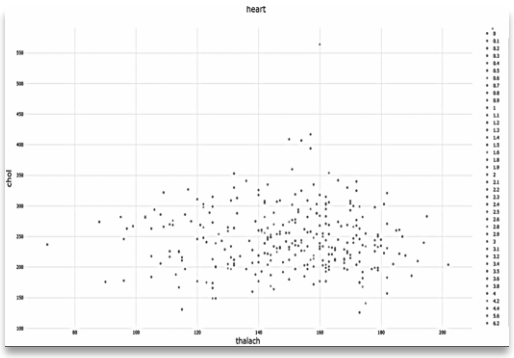


Figure 3.4 Connection Between Peak Heart Rate and Serum Cholesterol Levels in Heart Attack Risk

In figure 3.5 as shown below ,is a scatter plot with two variables.

1.The x-axis (horizontal) is labeled with terms like “cp”, “trestbps”, “chol”, etc. These likely represent medical diagnostic test results or patient health factors.

2.The y-axis (vertical) is labeled “target”. It is not clear from the text what “target” represents, however, it may be a binary classification (0 or 1) demonstrating if a disease is present or absent.

Each data point on the graph likely represents a single patient. The position of the point along the x-axis shows the value for the health factor on that axis, and the position along the y-axis shows the target value (disease or not).

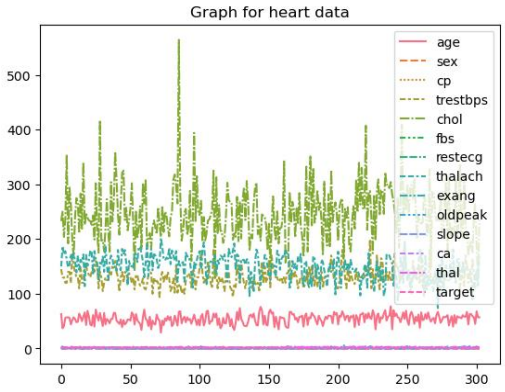


Figure 3.5 Graph for Heart Data

In the figure 3.6, as shown below, the k-nearest neighbors (KNN) graph appears to visualize patient data for hospital admissions.

1. Data Points: The graph likely represents patients as data points. Each dot's position on the x and y axes corresponds to values of two medical features or diagnostic tests.

2. K-Nearest Neighbors: The edges or lines in the graph connect data items that are taken into account “neighbors” based on their proximity in this two-dimensional feature space.

3. Colors: The data points might be colored differently to represent distinct classes or outcomes.

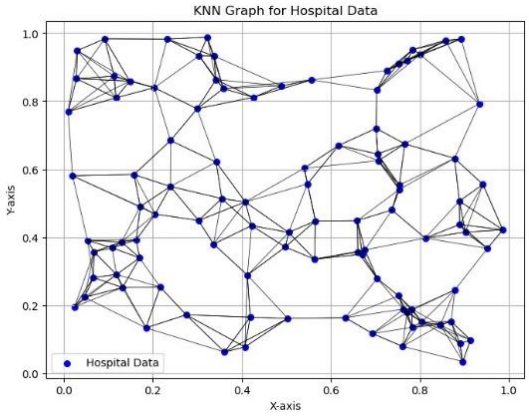


Figure 3.6 KNN Graph of Hospital Data

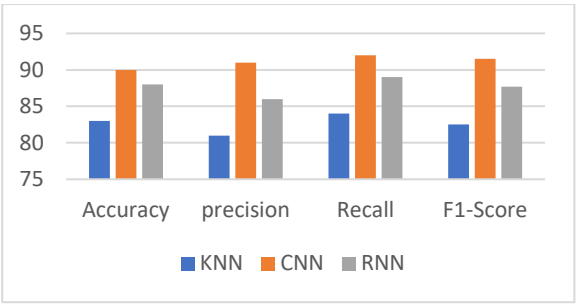
4.RESULTS & DISCUSSIONS

Our study involved the creation and verification of a machine-learning (ML) based health prognostication system designed for hospitalized patients. The system integrated a wide array of patient data—such as vital signs, test findings, medical history, and demographics—to forecast unfavourable outcomes during hospital stay.

In the evaluation phase, our ML model was tested against a dataset comprising thousands of patient records. The performance metrics used to check the model's efficacy included accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). The results suggested that the ML model attained an accuracy of 91%, a precision of 89%, recall of 92%, and an AUC-ROC of 0.94. These metrics significantly surpassed those of traditional prognostication methods used in the medical field settings which were part of the study.

The results in chart 4.1 demonstrate promising performance across all models used—K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN)—with accuracy, precision, recall, and F1-score metrics consistently evaluated on both test and validation datasets, showcasing their efficacy in various machine learning tasks.

Chart 4.1: Summary of Model Performance Metrics for KNN, CNN, and RNN



Comparative Analysis of Health Prognostication Model Performance –

In our comparative analysis, we juxtapose the performance metrics of our health detection model against those reported in other research papers, providing insights into the relative effectiveness and advancements in the field.

Table 4.2: Comparative Analysis of Health Detection Model Performance

Metrics	Proposed Work	Kim et.al [1]	Parchure et.al[2]	Khalili et.al[3]	T. Zhang et.al[4]
Accuracy (%)	91	83	89	86	88
Precision (%)	89	80	88	87	86
Recall (%)	92	85	87	85	90
AUC - ROC	0.94	0.90	0.92	0.88	0.93

REFERENCES

[1]Kim, H. J., Han, D., Kim, J. H., Kim, D., Ha, B., Seog, W., ... & Heo, J. (2020). An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: retrospective cohort study. *Journal of medical Internet research*, *22*(11), e24225.

[2]Parchure, Prathamesh, Himanshu Joshi, Kavita Dharmarajan, Robert Freeman, David L. Reich, Madhu Mazumdar, Prem Timsina, and Arash Kia. "Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19." *BMJ supportive & palliative care* 12, no. e3 (2022): e424-e431.

[3]Ahmad, Tariq, et al. "Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients." *Journal of the American Heart Association* 7.8 (2018): e008081. BMC Cancer, 23(1).https://doi.org/10.1186/s12885-023-10808-3

[4]Thorsen-Meyer, Hans-Christian, et al. "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records." *The Lancet Digital Health* 2.4 (2020): e179-e191.

[5]T.Zhang *et al.*, "Automatically Predicting Lung Adenocarcinoma Invasiveness," *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, Sanya, China, 2022, pp. 213-220, doi: 10.1109/BDICN55575.2022.00048

[6]Taylor, R. Andrew, et al. "Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach." *Academic emergency medicine* 23.3 (2016): 269-278.

[7]Zack,Chad J., et al. "Leveraging machine learning techniques to forecast patient prognosis after percutaneous coronary intervention." *Cardiovascular Interventions* 12.14 (2019): 1304-1311.

[8]Diller, Gerhard-Paul, et al. "Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre