Clinical Outcome Prediction Using K-Nearest Neighbors Algorithm

Aniruddha Prabhu B.P.[1], Aravindh P[2], Ashvin Vincent[3], Chayan Nath[4], Gopika N[5]

Dept. of Computer Science and Engineering, Cambridge Institute of Technology, Bangalore, India.

{aniprabhubp, aravindh.p201.741, ashvinvince , chayannath19, gopikakv1}@gmail.com

---***---

Abstract - Forecasts or suggestions supplied by data-driven computers are now influencing clinical decision-making in health care. In the recent clinical literature, several machine learning implementations have been developed, particularly for outcome prediction models, with outcomes ranging from death and cardiac arrest to acute renal damage and arrhythmia. To make predictions, machine learning (ML) discovers patterns in enormous volumes of medical data. The test person's symptoms are given into our Machine Learning (ML) model, and the disease is predicted as the algorithm's output. Our paper attempts to improve medicine and clinical practice by identifying patterns in massive volumes of medical data and making future predictions. Preventative measures are also included in the model to keep the patient's health from worsening.

Key Words: Machine Learning; Prediction;

1. INTRODUCTION

Disease prediction utilizing different data modalities, including voice signals and medical imaging and clinical outcome prediction to identify deterioration, such as cardiac arrest and death, have all proven effective with ML applications. Routinely, clinicians must assess multiple sources of information when determining the best course of treatment for their patients. Computer-generated forecasts sometimes outperform the predictions of tenured, experienced clinicians. With huge volume of dataset, the dataset represents a rich source of healthcare data for machine learning algorithms to make useful predictions. Also known as clinical decision support systems (CDSS), such data-driven tools are able to establish a short-term baseline from which clinicians can improve clinical outcomes by optimizing patient care. Prediction models are intended to aid healthcare practitioners and patients in making decisions regarding medical tests, medication initiation and discontinuation, and behavioral changes. Although they cannot replace professional expertise, they may offer reliable information about a person's disease risk and help to avoid certain typical prejudices in clinical decision-making. Predictive modelling using integrated Electronic Health Records (EHRs) is intended to increase patient delivery, reduce waste, and extend clinical expertise all at the same time [1]. Patient profiles are used in risk assessment models to quantify the likelihood that a certain outcome is present or may occur. The model also includes precautionary steps to prevent the patient's condition from deteriorating further [3]. We investigated the key challenges facing implementation of machine learning models into clinical practice, from the level of model development to evaluation. Currently, there is a lack of standardization, including the timing and frequency of update for data-driven computer forecasts that inform patient care. We present one such approach (a supervised K-Nearest Neighbor regression algorithm), discuss its key strengths and weaknesses, and outline recommendations for best practices in developing and evaluating data-driven algorithms.

2. LITERATURE REVIEW

A. Artificial intelligence in healthcare, 2018.

Artificial intelligence (AI) is influencing medical practice. Because of recent advancements in digital data collection, machine learning, and computer infrastructure, AI applications are expanding into fields that were previously thought to be solely the province of human knowledge [2].

B. A clinically applicable approach to the continuous prediction of future acute kidney injury. Nature, 2019.

Failure to recognize and treat deteriorating patients accounts for roughly 11% of hospital mortality, therefore early identification of deterioration might be beneficial in aiding healthcare providers. To achieve this aim, patient risk predictions must be updated and accurate regularly and provided to individuals with sufficient context and time to respond [6]. We utilize acute kidney injury as an example to build a deep learning method for predicting future deterioration in patients, relying on previous work that model's adverse events from electronic health records and utilizing it as a model.

C. Risk Prediction with Electronic Health Records: A Deep Learning Approach.

In recent years, there has been a surge in interest in data analytics due to patient Electronic Health Records. One of the most promising approaches for revolutionizing healthcare is data-driven healthcare, which promises to make good use of large medical data, which represents the knowledge gathered from the treatment given to hundreds of millions of patients, to provide the best customized treatment possible. Working directly with medical datasets has numerous obstacles, including temporality, sparsity, noisiness, bias, and so on [7]. As a result, extracting useful features or phenotyping from patient Electronic Health Record's is a crucial step before moving on to other applications. The suggested approach is tested on a data warehouse from the real-world Electronic Health Record in the context of chronic illness prediction modelling.

3. FRAMEWORK OF CLINICAL OUTCOME PREDICTION MODEL

Due to the extreme variety of admitted patients, care paths within hospitals vary greatly. As a result, understanding the clinical environment is critical for designing machine learning models that can be integrated into existing medical processes. Early warning systems and other patient monitoring approaches are extensively utilized across hospital wards to screen for medical problems regularly.

A. Methodology

Our Machine learning model identifies trends in vast volumes of medical data that are datasets we have chosen which contain columns including symptoms (variables) and corresponding diseases (target function). We can make potential predictions of the diseases present in the test subject. The input to our Machine Learning (ML) model is the symptoms exhibited by the test subject, whose disease is then predicted as the output of the model. Data transformation entails converting the categorical values we have to numerical values to process the data. Then we divide values in the datasets into two sets, a training set and a test set and perform preprocessing using the Feature Scaling approach [5]. Processed data is a dataset containing processed data to train the model. We train the model using the K- Nearest Neighbors algorithm [4]. User information is given to the system. The user will log in or register (if they are a first-time user) and fill out their information. User input entails the user entering their symptoms or the symptoms faced by a test subject. These inputs will be supplied into the disease prediction model that has been trained. The outcome is the model will forecast the disease as an output based on the symptoms provided as input.

B. Pre-processing Data

- **Handling missing values** - Missing values are undesirable, but quantifying the magnitude of effects in machine learning models is challenging. Missing data, in general, causes imbalanced findings, skewed calculations, and, in extreme cases, invalid predictions. The missing values in the dataset are filled up with a common value to thus avoid these issues from occurring.
- **Label encoding** - Label encoding is the process of turning labels into numeric representations so that machines can read them. Then, using machine learning algorithms, informed decisions about how to use those marks will be made. In supervised learning, it is a crucial pre-processing phase for the structured dataset.
- **Feature scaling** - The characteristics will be rescaled to resemble a typical normal distribution with standardization (or Z-score normalization). The following is how the sample standard scores (also known as z - scores) are calculated:

$$z = \frac{x - \mu}{\sigma}$$

where σ is the standard deviation from the mean and μ is the mean (average).

C. K – Nearest Neighbors

K-Nearest Neighbors is a simple, straightforward supervised machine learning approach that may be used to solve classification and regression problems. Similar things are close to each other, according to the KNN technique. To put it another way, close objects are linked together.

1. Load the data.
2. Change K to the desired number of neighbors.
3. For every instance in the data
   3.1. Calculate the distance between the query and current data samples.
   3.2. Add the distance and the sample's index to an ordered collection.
4. From smallest to greatest, sort the ordered collection of distances and indices (in ascending order).
5. From the sorted list, choose the first K entries.
6. Get the labels for the K entries you've chosen
7. Return the mean of the K labels on the condition that a regression is found.
8. Once classification is done yield the mode of the K labels.

Choosing the appropriate K value for implementation: To find the optimal number of nearest neighbors (K) for our results, we run the K-Nearest Neighbors algorithm several times with different values for K to find the one that reduces the number of errors while still allowing the system to make right predictions.

Euclidean Distance: To compute the interval between two existing data points we use the Euclidean distance formula given by –

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

where, d is the distance between two points in the plane with coordinates (X1, X2) and (Y1, Y2).

D. Implementation

For training the Machine Learning model we used two data files. One containing details about the diseases and their symptoms and the other file containing details about how much each symptom would contribute towards the prediction of the presence of a particular disease. Before training the model, we process the data by filling in the places in the dataset which had missing values. While training any Machine Learning model we need to replace missing values. Generally, we tend to replace the missing values with mean values. In our model we prefer filling the missing values with zeroes instead of mean values, to avoid any confusion that the mean values would cause in the prediction of a certain disease. As our severity data file also contains numerical values replacing the missing values with median values could misguide the model in predicting diseases that the test subject might not have or might have in the future. In the symptoms and diseases dataset, we replace the categorical values of the symptoms into numerical values using the severity

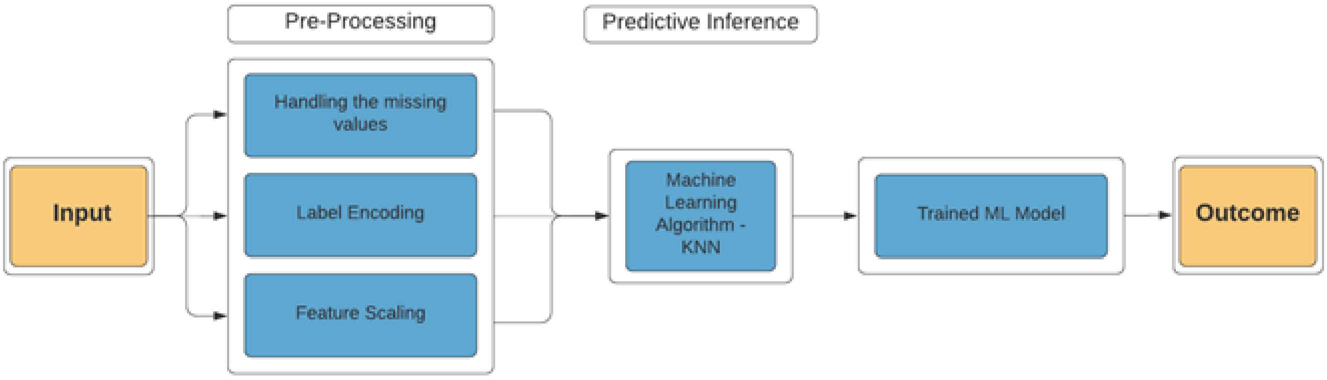

Fig 1: A general machine learning pipeline that predcits the outcome from the input. The two main stages are Pre-processing and training the model.

dataset. Using Label Encoder, we replace all the categorical values of the diseases with numerical values. A test set and a training set were created from the dataset. We utilize 80% of the original dataset to train the model and 20% of the original dataset to test it. Then we perform Feature Scaling [5], which involves the Standard Scaler method. We ensure that regardless of the range of difference existing between the larger or smaller values the model would give equal preference to all values while predicting the disease present in the test subject. This completes our pre-processing of data. We then train the model with the training set of the data. Now, we use the test set to predict the diseases that the test subject could have. We take all the predicted values and test set values which are currently in numerical form convert them back to their original form which was categorical values. We then concatenate the test set values and predicted values of the diseases into a 2-D matrix. We then generate a result file containing the values predicted by the model.

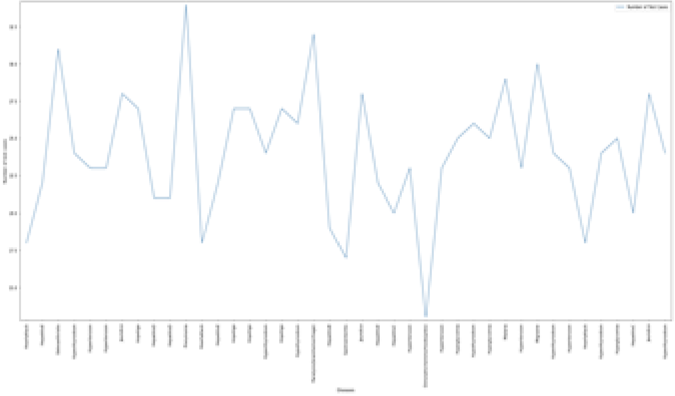

Fig 2: The graph of number of test cases vs. diseases.

4. PERFORMANCE EVALUATION

When the output can be divided into two or more categories, the easiest way for solving a classification problem is to use performance measures. A table with two dimensions: real and predicted, as well as True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) on both dimensions, is known as an uncertainty matrix.

A. Accuracy of Classification

For classification algorithms, it is the most commonly used performance statistic. The share of correctly classified data instances among all data instances. Using an uncertainty matrix and the formula below, we can compute it simply.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

To calculate the accuracy of our classification model, we can utilize the accuracy_score function in sklearn.metrics.

B. Precision

Precision in document retrievals may be given as the fraction of retrieved instances that are relevant. It determines the relevance of the results. We can simply quantify it using a confusion matrix using the following formula.

$$Precision = \frac{TP}{TP + FP}$$

C. Recall

The fraction of relevant instances that were retrieved by our ML model is known as recall. With the aid of the following formula, we can easily measure it using a confusion matrix.

$$Recall = \frac{TP}{TP + FN}$$

5. RESULT

In this part, we examine the experimental data acquired to verify the efficacy of the suggested method. We evaluated the suggested model on 984 test instances to determine its robustness, and the related quantitative metrics were produced. The clinical outcome prediction algorithm properly predicts 971 test cases out of 984. The quantitative measures (accuracy, precision, and recall) have been derived based on this information and are listed in a table.

TABLE I: The quantitative measurements of the proposed approach obtained from the testing phase.

Parameters	Values
Accuracy	0.9867
Precision	0.99
Recall	0.99

6. CONCLUSION

The suggested architecture demonstrates that any healthcare company may use machine learning to save expenses and serve patients. It simplifies the decision-making process in terms of illness early detection and diagnosis. It not only processes the patient's raw data or symptoms, but also provides reports promptly, allowing researchers, physicians, nurses, and other stakeholders to receive the assistance they require. This paper also highlights breakthroughs in the field of machine learning in healthcare that may be included in the suggested architecture. To put it another way, the suggested approach would improve healthcare professionals' ability to comprehend their patients' fundamental requirements, allowing them to more effectively counsel and support their patients while also maximizing staff time and lowering expenses. Healthcare professionals and hospitals all over the world must adopt ML-based systems because they offer a slew of benefits to the industry.

7. REFERENCES

[1] Machine Learning for Clinical Outcome Prediction Farah Shamout, Tingting Zhu, and David A. Clifton.

[2] Kun Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. Artificial intelligence in healthcare, 2018.

[3] Farah E. Shamout, Tingting Zhu, Pulkit Sharma, Peter J. Watkinson, and David A. Clifton. Deep Interpretable Early Warning System for the Detection of Clinical Deterioration. IEEE Journal of Biomedical and Health Informatics, 2019.

[4] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. International Journal of Information and Education Technology, 2(3), 220. Retrieved July 11, 2017.

[5] Hayden Wimmer and Loreen Powell. Principle Component Analysis for Feature Reduction and Data Preprocessing inData Science. In Proceedings of the Conference on Information Systems Applied Research, pages 1–6, 2016.

[6] Enad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O. Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R. Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R. Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767):116–119, 2019.

[7] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2016., pages 432–440, 2016.