Machine Learning-Powered Clinical Predictions: from Data to Deployment

Ayush Nautiyal, Sagar Negi, Harshit Bajpai and Chinmay Raj Shah

May 31, 2024

Machine Learning-Powered Clinical Predictions: From Data to Deployment

Ayush Nautiyal
Graphic Era University
Dehradun, India
anautiyal3355@gmail.com

Sagar Negi
Graphic Era University
Dehradun, India
negis2673@gmail.com

Harshit Bajpai
Graphic Era University
Dehradun, India
harshit16bajpai@gmail.com

Chinmay Raj Shah
Graphic Era University
Dehradun, India
shah.chinmayraj.2016711@gmail.com

***Abstract*—The integration of sophisticated machine learning algorithms into clinical applications has the potential to transform healthcare by providing highly accurate predictive models. This case study focuses on the design, development, and evaluation of clinical predictive applications, with a primary emphasis on machine learning methodologies. The article begins by elucidating the motivation for this initiative, emphasizing the urgent need for advanced predictive models to improve healthcare outcomes. The architectural design and implementation of the application are discussed, highlighting the central role of machine learning at each stage.**

The study details the comprehensive integration of machine learning algorithms, covering crucial aspects such as data preprocessing, feature extraction, model training, validation, and deployment. Various machine learning techniques, including classification, regression, and clustering, are rigorously analyzed for their effectiveness in predicting clinical outcomes, with a specific focus on pain prediction. The study examines the performance of different models and their respective algorithms, providing a detailed comparison to determine the most effective approaches.

Challenges encountered during the development process, such as handling preliminary data, selecting appropriate models, and optimizing algorithms, are explored in depth. Strategies to mitigate these challenges are discussed, along with their impact on the application's predictive performance. The paper concludes with a thorough discussion of the research outcomes, highlighting the significant advantages, potential limitations, and future research directions in the application of machine learning to clinical prediction.

This work underscores the transformative power of machine learning algorithms in developing robust, scalable, and highly accurate medical applications, demonstrating a substantial advancement in healthcare predictive capabilities.

I. Introduction

Providing the best care to all patients is a great challenge in medicine or healthcare, and only those who can afford it can benefit from it. Today, the health sector has become a large and profitable sector. It can also be difficult for users to reach doctors and hospitals due to the disease not being diagnosed. Therefore, it would be better for the patient if the above process were done using automated software, which saves time and money and makes the process smoother. In this digital world, information is an asset, and big data is being created everywhere. Data obtained from the medical sector includes all information about the patient. This medical information will also be used to provide effective and efficient treatment to healthy patients. The field also needs some improvements in the use of information from health research. However, extracting information from data is a major challenge. This requires some data mining and machine learning techniques due to the large amount of data. A wealth of medical information is available, but it is not up to the large amount of data reliably researched to reveal the hidden knowledge necessary for successful decision-making. Machine learning technology provides healthcare with a powerful platform to solve health problems effectively. Your relative or loved one may need to urgently seek help from a doctor for something important but cannot get a doctor's advice due to a prior agreement or other obvious reasons. That's when this automated service comes into play. Disease Predictor is a web-based program that predicts a user's illness based on their symptoms. Disease prediction draws information from various health-related websites. The necessity of this study is to prevent the risk of premature death by predicting diseases, saving people's lives, and reducing medical costs to some extent. Here, we use some smart data processing techniques to get the most accurate patient details about the disease. We will use machine learning algorithms to obtain a multi-path prediction. Python pickling is used to save the model. The importance of this analysis is that when identifying the disease, it includes all the bacteria that cause the disease, so the disease can be diagnosed effectively and accurately. The final model behavior will be saved as a Python pickle archive. This system is useful for people to check it every day, allowing people to understand their health and encouraging them to adopt healthy habits. According to research, such systems are not widely used, and little is known about them. Following this principle can help people avoid unnecessary visits to the hospital by using this free app wherever they are.

II. Literature Review

AIDaroos KM [1] Research review and information about the best and easiest refining mining methods. Completed. forever. The authors compared Nivea with five other tax systems. Its basis is logistic regression (LR), Kstar (K*), decision tree (DT), neural network (NN) and simple rulebased algorithm . It uses 15 realworld medical problems from the UCI Machine Learning Repository . People are selected for research. Package 8 failed to find 8 of the 15 datasets that were considered to be better prediction methods than others to tackle. The results show that the tree works very well and sometimes is at least as accurate as: Decision tree in Bayesian taxonomy.good. It is also proposed to use genetic algorithms to reduce the amount of data to obtain the symptoms necessary to evaluate heart disease, thus further improving the accuracy of decision trees and Bayesian classification. You compare exercise values and then check heart rate. The results of Masathe H.D and colleagues, showing 99Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-Laszlo Barabasi have shown that treating infectious diseases worldwide does not save time or money. Therefore, the authors conducted this study to predict possible diseases. CARE (estimating the risk of infection using only the patient's medical history and ICD-9-CM codes) was used for this purpose. CARE provides an integrated approach to predicting each patient's greatest risk of infection based on his or her medical history and similar patient populations. The authors also describe ICARE, an iterative process involving integrated models to improve performance. These state-of-the-art machines do not require any experience and can predict many diseases in a single operation. ICARE's superior coverage means earlier warning for thousands of diseases that could be reported years in advance. When fully implemented, the CARE system can be used to investigate the broader context of disease, raise previously unanticipated questions, and facilitate discussion of early prevention research and practice. As discussed in the introduction, some research articles include various models used to predict the disease a patient may have based on the symptoms the patient records. Currently, the most used and accurate model is as follows: the method proposed by Jianfang et al. used vector machines (SVM) to classify diseases based on symptoms. The SVM model is effective for disease prediction, but more time is needed to predict the disease .Moreover, the method cannot improve the accuracy of the model. The disadvantage of this method is that hyperplanes are used to divide objects, which can only be used in certain ways . Hyperplanes are only accurate when separating sample data into 2 groups. However, currently the medical industry needs more than 2 groups (diseases) to describe the symptoms associated with the disease. The K-most neighbors (KNN) algorithm used by Keniya et al. They use this method to be sensitive to noise and missing data by assigning data points to the cluster where most of the K data points are located. They take into account certain factors such as age group, symptoms and gender to predict the disease. When these parameters are taken into account, machine learning models become less accurate. The KNN method was also used by Kashvi et al. They have also been shown to be more accurate in many diseases, including diabetes and heart disease. There is a problem of determining small data in the classification of diseases . Pinale et al. The method suggested by. They predicted rare diseases such as diabetes, malaria, jaundice, dengue fever, and tuberculosis using the negative Bayes method. They have not yet worked on big data to predict many diseases . Additionally, Gomathy and Rohith Naidu used the negative Bayes method for disease prediction. Using this approach, they created a web-based virus prediction application that can be accessed from anywhere. The accuracy of the model depends on the data provided to the system. The challenge of the proposed model is to develop disease prediction software with more accurate data to improve accuracy . The method proposed by Chhogyal and Nayak uses the Naive Bayes classifier. Their accuracy in disease prediction is poor and they are not trained using standard data . The method proposed by Kumar et al. used the Rustboost algorithm. RUSSBoost was developed to solve the underclassification problem . However, the RUSBoost algorithm uses random subsampling as an iterative method, which may result in loss of important information. Therefore, the algorithm is not taken into account when learning the data. The above methods discuss various machine learning methods for disease prediction. However, the authors did not take advantage of issues such as efficiency, accuracy, limited data used to inform the model, and consideration of a limited number of symptoms to diagnose the disease. In order to overcome all these problems, a modified and accurate model that can predict human diseases should be prepared. The detailed model is described in the following section.

III. Methodology

A. Creation of the dataset

The dataset has been referenced from the website medlineplus, which is the government site of the United Kingdom (UK), and all the diseases mentioned have been verified and cited by the specialists (link attached in reference section).

For the test data, the symptoms are derived from different sources, which are mentioned in the reference section.

B. Model training

The dataset is trained and tested on 11 models which are used for classification tasks.

1) Logistic Regression: Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between zero and one. This function is known as the logistic function.

$$p(x;b,w) = \frac{e^{\beta_0 + x\cdot\beta}}{1 + e^{\beta_0 + x\cdot\beta}} = \frac{1}{1 + e^{-(\beta_0 + x\cdot\beta)}}$$

Fig. 1. Logistic Regression Equation

2) Extra Tree Classifier: Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple decorrelated decision trees collected in a "forest" to output it's classification result.

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i)$$

Fig. 2. Entropy Equation

$$Gain(S,A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Fig. 3. Information Gain Equation

3) Naive Bayes: Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

Fig. 4. Bayesian Equation

4) Random Forest Classifier: Random Forest Classifier is an ensemble learning method using multiple decision trees for classification tasks, improving accuracy. It excels in handling complex data, mitigating overfitting, and providing robust predictions with feature importance.

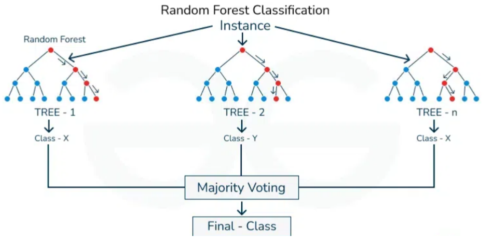


Fig. 5. Random Forest Classifier

5) SVM: A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.

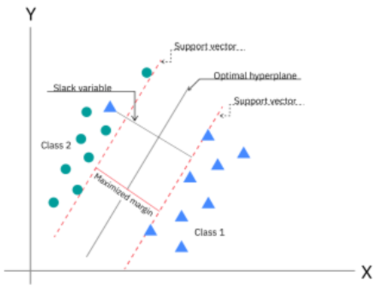


Fig. 6. SVM hyperplane

6) Ridge Classifier: The Ridge Classifier is a machine learning algorithm designed for classification tasks. By combining ideas from conventional classification techniques and Ridge Regression, it offers a distinct method for classifying data points. The L2 regularization used by the Ridge Classifier, which has its roots in Ridge Regression, stops overfitting by adding a penalty term that is managed by the hyperparameter alpha. This regularization aids in preserving equilibrium between managing model complexity and fitting the data. Its ability to adapt classification to a regression framework by transforming target variables into a specified range, usually between -1 and 1, is one of its distinguishing features. This conversion reduces the chance of overfitting.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Fig. 7. L2 Regularization

7) Decision Tree Classifier: Decision Tree is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees.

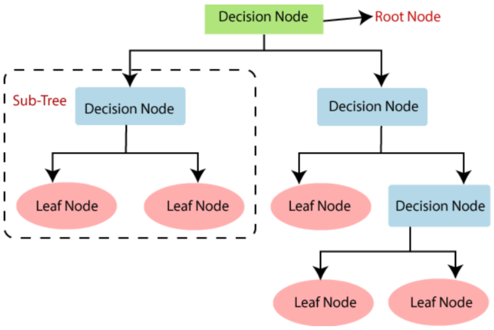


Fig. 8. Decision Tree

8) Gradient Boosting Classifier: Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

y(pred) = y1 + (eta * r1) + (eta * r2) + ... + (eta * rN)

9) K Neighbors Classifier: The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.For classification problems, a class label is assigned on the basis of a majority vote that is the label that is most frequently represented around a given data point is used.

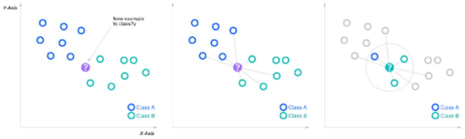


Fig. 9. KNN

10) Ada Boost Classifier: Adaptive Boosting is an ensemble learning used in machine learning. The main idea behind AdaBoost is to iteratively train the weak classifier on the training dataset with each successive classifier giving more weightage to the data points that are misclassified.

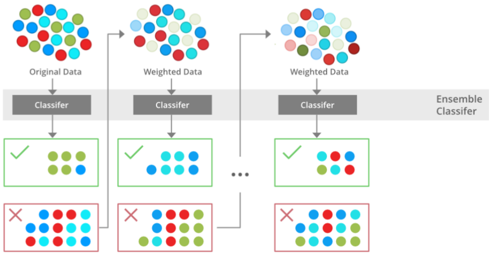


Fig. 10. Ada Boost Classifier

11) XG Boost Classifier: XGBoost, or eXtreme Gradient Boosting is a machine learning algorithm that belongs to the ensemble learning category, specifically the gradient boosting framework. It utilizes decision trees as base learners and employs regularization techniques to enhance model generalization.

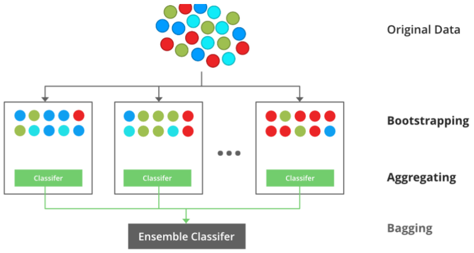


Fig. 11. XG Boost Classifier

C. Frontend Development

React (also known as React.js or ReactJS) is a free, open-source JavaScript library for building object-oriented user interfaces. It is managed by Meta (formerly Facebook) and a community of developers and companies.

React can be used to create single-page, mobile or server-rendered applications using frameworks such as Next.js. Since React only deals with the UI and exports objects to the DOM, React applications often rely on libraries for rendering and other client-side operations. The main advantage of React is that it only re-renders the parts of the page that have changed, thus avoiding unnecessary re-refreshing of DOM elements that have not changed. React allows you to create user interfaces using separate components called widgets. Create your own React elements like thumbnails, like Buttons, and videos. Then place them across all screens, pages, and apps. Create a dialog to get symptoms and display prediction values using Reactjs.

D. Backend Development

The backend is the code that runs on the server, receives requests from the client, and contains the logic to send the appropriate information back to the client. The backend also includes a database that will store all the data for the application. This article focuses on the server-side hardware and software that make this possible. It is the engine behind the scenes that ensures a good user experience, secure data storage, and user loyalty. The database uses MongoDB and Node.js. Node.js is a cross-platform, open-source JavaScript runtime environment that runs on Windows, Linux, Unix, macOS, and more. Node.js runs on the V8 JavaScript engine and runs JavaScript code outside of the web browser.

Node.js has an event-driven architecture that can perform asynchronous input and output. The purpose of these design choices is to optimize deployment and optimization capabilities for real-world web applications as well as multiple input and output web applications.

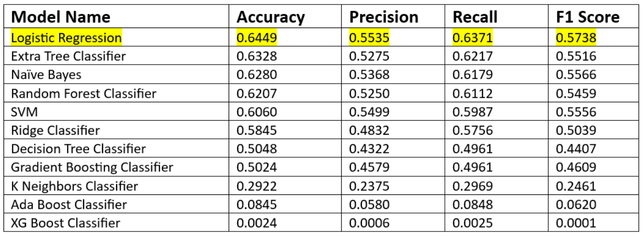
IV. Results

A. Dataset

The dataset is created successfully and contains 391 disease with 1330 symptoms.

B. Model Selection

By comparing different models on the performance metrics the Logistic Regression comes to be best with accuracy of 64.49



Model Name	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6449	0.5535	0.6371	0.5738
Extra Tree Classifier	0.6328	0.5275	0.6217	0.5516
Naïve Bayes	0.6280	0.5368	0.6179	0.5566
Random Forest Classifier	0.6207	0.5250	0.6112	0.5459
SVM	0.6060	0.5499	0.5987	0.5556
Ridge Classifier	0.5845	0.4832	0.5756	0.5039
Decision Tree Classifier	0.5048	0.4322	0.4961	0.4407
Gradient Boosting Classifier	0.5024	0.4579	0.4961	0.4609
K Neighbors Classifier	0.2922	0.2375	0.2969	0.2461
Ada Boost Classifier	0.0845	0.0580	0.0848	0.0620
XG Boost Classifier	0.0024	0.0006	0.0025	0.0001

Fig. 12. Comparison of models

When confusion matrix is plotted between true values and predicted values then mostly result predicted is true positive.

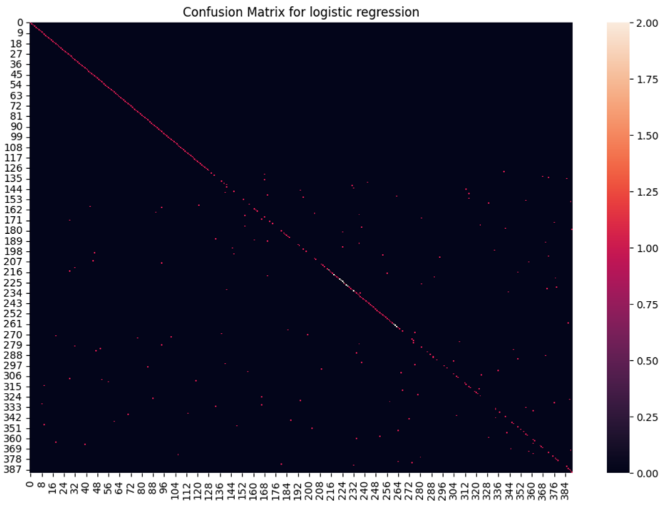
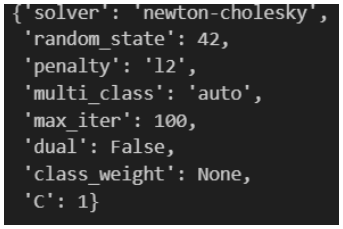


Fig. 13. Confusion Matrix of Logistic Regression

After hyper-tuning parameters the best estimators are achieved.

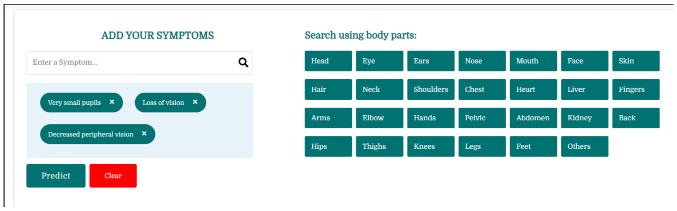


```
{'solver': 'newton-cholesky',
 'random_state': 42,
 'penalty': 'l2',
 'multi_class': 'auto',
 'max_iter': 100,
 'dual': False,
 'class_weight': None,
 'C': 1}
```

Fig. 14. Parameters after hypertuning

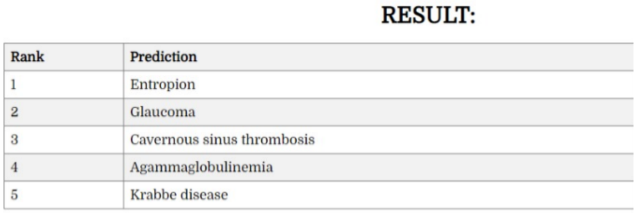
C. Website Development

Users can enter the details by selecting the details of the symptoms and for the users convenience each symptom is categorized according to the body parts and top five predicted results sorted in descending order is shown to the user.



Very small pupils ×
Loss of vision ×
Decreased peripheral vision ×

Fig. 15. Symptoms Selection



RESULT:

Rank	Prediction
1	Entropion
2	Glaucoma
3	Cavernous sinus thrombosis
4	Agammaglobulinemia
5	Krabbe disease

Fig. 16. Result after symptoms selection by user

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our mentor, Mr. Priyank Pandey, Assistant Professor, Graphic Era University, for his continuous support, patience, and immense knowledge throughout our project development. His guidance helped us in all the time of development and writing of this paper.

We are also grateful to the Department of CSE at Graphic Era University for providing the necessary facilities and resources. We are profoundly grateful to our families and friends for their unwavering support and encouragement. To our parents, who instilled in us, the value of education and hard work.

Finally, thanks to all the individuals who have contributed to this research directly or indirectly.

REFERENCES

[1] Symptoms and diseases of dataset referenced from website medlineplus "https://medlineplus.gov/encyclopedia.html“.

[2] Sneha Grampurohit and Chetan Sagarnal, "Disease Prediction using Machine Learning Algorithms“,2020 International Conference for Emerging Technology (INCET), Belgaum, India, Jun 5-7, 2020.

[3] Kriti Gandhi, Mansi Mittal, Neha Gupta and Shafali Dhall, "Disease Prediction using Machine Learning“, International Journal for Research in Applied Science and Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429, Volume 8 Issue VI June 2020, Available at "www.ijraset.com“.

[4] Er. Harjeet Singh and Mr. Ankit Mehta, "DISEASE PREDICTION SYSTEM USING SYMPTOMS“, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 10 Issue: 06 — Jun 2023, Available at "www.irjet.net“.

[5] Divya Mandem and B. Prajna, "Multi Disease Prediction System“,IJIRT, Volume 8 Issue 6 , ISSN: 2349-6002, November 2021.

[6] Ankush Singh, Ashish Yadav, Saloni Shah and Prof. Renuka Nagpure, "Multiple Disease Prediction System“, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 09 , Issue: 03 — Mar 2022, Available at "www.irjet.net“.

[7] Gaurav Shilimkar, Gaurav Shilimkar and Shivam Pisal, "Disease Prediction Using Machine Learning“, International Journal of Scientific Research in Science and Technology, Print ISSN: 2395-6011 — Online ISSN: 2395-602X (www.ijsrst.com)