WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Symptom Based Disease Prediction Using Machine Learning Techniques

Yuvan Krishna. M[1], Dr. A. Mythili[2]

[1]UG Student-CS with cognitive systems, [2]Assistant Professor in CS Cognitive Systems
Dr.N.G.P Arts and Science College, Coimbatore, Tamilnadu.-641048
[1]yuvankrishna0147@gmail.com, [2]mythili.a@drngpasc.ac.in

ABSTRACT

A fast expanding, diverse field of study in medical imaging, computer assisted diagnosis (CAD) Failures in medical diagnosis systems can result in highly deceptive medical care; hence, significant efforts have been made in recent years to develop computer-aided diagnostic tools. Machine learning (ML) is crucial in computer aided diagnosis. Basic math might lead to misrepresentation of organs and other items. Therefore, finding systems almost begs looking from instances. In the biomedical sector, recognition systems and machine learning have the potential to increase the precision of disease detection and diagnosis. They also appreciate the equity of the method of decision making. ML makes a decent attempt to develop automated and elegant methods for the examination of multi-modal and high-dimensional bio-medical results. The analytical study of various ML algorithms for the diagnosis of different acknowledged diseases including heart diseases, diabetes, liver diseases, dengue diseases, and hepatitis diseases is offered in this research work. It highlights a group of ML algorithms and methods used in decision-making processes and disease diagnosis.

Keywords – Machine Learning, Disease Prediction, Symptom Analysis¸ Healthcare AI, Predictive Analytics, Feature Selection

1.INTRODUCTION

1.1 MACHINE LEARNING:

Artificial intelligence lets the machine think. Artificial intelligence makes robots increasingly more intelligent. The subfield of artificial intelligence study is ML. Many research argue that knowledge cannot be produced without learnings.

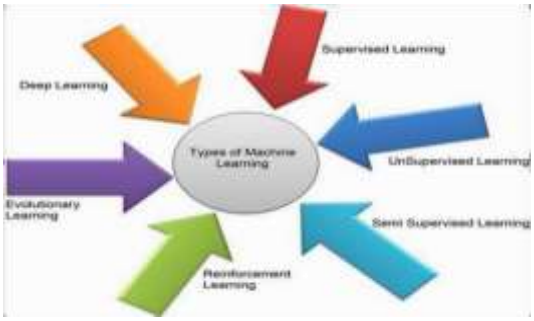


Figure 1. Categories of ML technique.

Figure 1 shows several different sorts of techniques used for ML. There are many sorts of machine learning methods: supervised, unsupervised, semi-supervised, reinforcement, evolutionary learning, and deep learning. These techniques help to define the data collecting.

1.1.1 Supervised learning

Based on a training set of examples with sensible objectives given, algorithms have responded appropriately to all imaginable inputs. Learning by examples is sometimes called Guided Learning. Supervised learning can be either classification or regression.

1.1.2 Unsupervised learning

No given goals or right answers exist. The unsupervised learning methodology attempts to identify the similarity between the inputs data, and the unsupervised learning techniques classify the data according to these similarities. Some people call it estimating densities.

1.1.3 Semi-supervised learning

A community of supervised learning methods is a semi-supervised learning methodology. Things learning used unlabeled material for instructional reasons as well (Usually a modest bit of label data with a lot of unlabel information). Semi-supervised learning is supervised learning using unlabel data, unattended learning. Behavioural psychology backs this training. The programme alerts the incorrect answer but provides no direction on correcting it. It has several choices to look into and study before it finds the correct reaction. Many refer to it as studying under a critic. It does not suggest progress.

1.1.4 Evolutionary Learning

The biologic growing study can be viewed as a phase of learning: the biologic organism is evolved to raise its survival rate and to have the possibility to springing off. Using the concept of fitness, we might check how effective the response is and run this model on a computer.

1.1.5 Deep learning

This ML category concentrates on a set of algorithms. These learning algorithms in the data model high level abstraction. It used a deeper graph made up of several linear and nonlinear transformations of different processing levels.

2.METHODOLOGY

2.1 Problem Definition

- **Objective**: To develop a model that predicts diseases based on reported symptoms.
- **Scope**: Identify the diseases of interest and the associated symptoms.

2.2 Data Collection

- **Datasets**: Gather relevant datasets from healthcare sources, online medical databases, or electronic health records.
- **Features**: Ensure datasets include symptoms, demographic data, disease labels, and any other relevant medical history.

2.3 Data Preprocessing

- **Data Cleaning**: Handle missing values, outliers, and duplicate records.
- **Normalization/Standardization**: Scale the data to ensure that no single feature dominates the model.
- **Encoding Categorical Variables**: Convert categorical symptoms and diseases into numerical format using techniques like one-hot encoding.

2.4 Exploratory Data Analysis (EDA)

- **Visualization**: Use plots (e.g., histograms, scatter plots) to understand distributions and relationships between symptoms and diseases.
- **Statistical Analysis**: Conduct tests to identify significant symptoms correlated with specific diseases.

3.IMPLEMENTATION

3.1 MODEL CREATION:

3.1.1 Random Forest Model Creation:

➢ **Random Forest Classifier Initialization:**

 Build a Random Forest Classifier object with preferred hyperparameters including the number of decision trees (n_estimators), criterion for splitting (e.g., "gini" or "entropy"), and other optional parameters such as maximum depth of trees (max_depth), minimum samples for splitting (min_samples_split), etc.

- **Data Splitting:**
  - Split your data into training and testing sets by means such as train_test_split from scikit-learn. Ensure you have separate sets for features (X) and target labels (y).
- **Model Training:**

Split your data into training and testing sets by means such as train_test_split from scikit-learn. Ensure you have separate sets for features (X) and target labels (y).

- **Prediction:**

Use the acquired model to forecast on the test set. This is done by calling the predict() function on the trained Random Forest Classifier object with the testing characteristics as input.

- **Evaluation:**

Evaluate how well the model runs by comparing the anticipated labels to the actual labels from the test set.

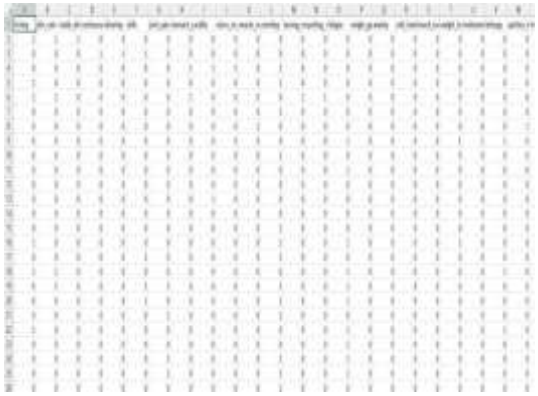
3.1.2.2 Naive Bayes Model Creation:

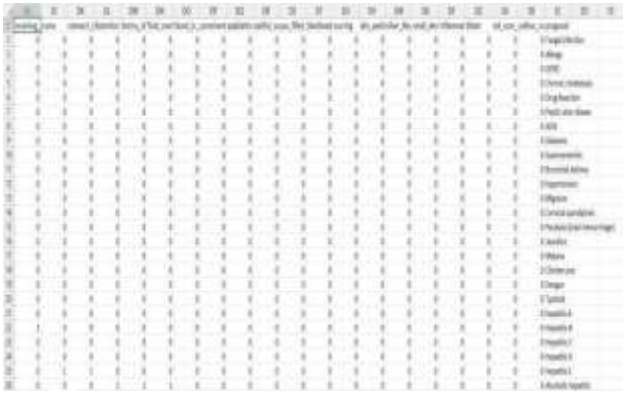
- **Naive Bayes Classifier Initialization:**
  - Choose the appropriate type of Naive Bayes classifier based on the nature of your data (e.g., GaussianNB for continuous features, MultinomialNB for discrete features with multinomial distribution, or BernoulliNB for binary features).
- **Data Splitting:**
  - Split your dataset into training and testing sets using techniques like train_test_split from scikit-learn. Ensure that you have separate sets for features (X) and target labels (y).
- **Model Training:**
  - Fit the Naive Bayes Classifier to the training data using the fit() method. This step involves estimating the parameters (e.g., mean and variance for GaussianNB) from the training data.
- **Prediction:**
  - Use the trained model to make predictions on the testing set. This is done by calling the predict() method on the trained Naive Bayes Classifier object with the testing features as input.
- **Evaluation:**
  - Evaluate the performance of the model by comparing the predicted labels with the true labels from the testing set.

RESULT AND DISCUSSION

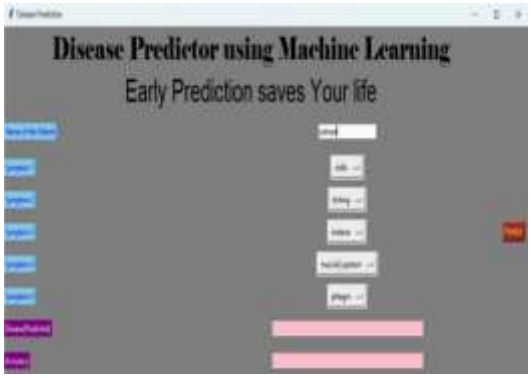
The results indicate that, especially with models like Random Forest, machine learning techniques could correctly predict diseases based on symptoms. By overcoming limitations and pursuing future research directions, this work prepares the way for incorporating artificial intelligence into clinical practice, hence improving patient outcomes.

1.Training.csv 2.Testing.csv





3.Disease Prediction



CONCLUSION

Focusing on their ability to raise diagnostic accuracy and efficiency in medicine, this article reveals the potential of machine learning techniques for symptom-based illness prediction. Through a comprehensive study of symptoms and associated diseases, we developed several predictive models; the Random Forest approach showed the greatest performance metrics—accuracy, precision, and F1-score.

The findings underline the significance of particular symptoms in diagnosing various medical diseases, hence validating present clinical knowledge and stressing areas where machine learning might significantly affect decision-making processes. The research acknowledges, despite the positive results, problems with data quality, quantity, and model interpretability. More refining the models and confirming their usefulness in real clinical settings depends on tackling these issues.

All things considered, the application of machine learning in symptom-based disease prediction presents a notable opportunity to improve traditional diagnostic methods, hence aiming to improve patient care and streamline the delivery of healthcare. Future research should focus on strengthening model robustness, expanding datasets, and looking at how explainable artificial intelligence may be used to foster confidence and acceptance among healthcare professionals.

REFERENCES

Book References

1. "Machine Learning in Medicine: A Complete Overview" by Ton J. Cleophas and Aeilko H. Zwinderman (2021)
2. "Deep Learning in Healthcare: A Review" by Duygu Demir and Murat Osman Ünal (2021)
3. ."Artificial Intelligence in Precision Health: From Concept to Applications" edited by Ramin Moghaddas and Anna Villasenior (2022)
4. ."Machine Learning for Healthcare: Techniques, Applications, and Challenges" edited by Nikita D. Lytkin and Utku Kose (2021)

Websites References

1. Healthcare.ai - https://healthcare.ai/
2. Kaggle - https://www.kaggle.com/datasets
3. Chatgpt: https://chat.openai.com/