Title 1: Disease Prediction using Machine Learning

Authors: Kriti Gandhi, Mansi Mittal, Neha Gupta, Shafali Dhall

Publication Date: June, 2020

Proposed System

The paper proposes a classification-based machine learning model for disease prediction. The system involves building and training various machine learning models on healthcare data to predict diseases based on symptoms and medical history. The authors emphasize the importance of early diagnosis, which can help in reducing the risk and improving treatment outcomes. The system utilizes feature selection techniques to enhance the efficiency of predictions.

Dataset used: The dataset comprises of 133 columns, comprising of 132 varied symptoms experienced by patients suffering from a range of ailments. A total of 40 diseases are present in this dataset.

Algorithms Used

The research utilizes several machine learning algorithms, including **K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Naïve Bayes, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Random Forest**. Each of these algorithms is tested for classification performance, with accuracy scores recorded for comparison.

Methodology for Increasing Accuracy

The authors use various feature selection techniques such as Recursive Feature Elimination (RFE) and embedded methods to refine the input data and improve prediction accuracy. They preprocess the dataset by handling missing values, performing feature selection, and dividing the data into training and testing sets. The study also evaluates different algorithms based on accuracy, processing speed, and efficiency to determine the most effective model.

Accuracy Generated

The results indicate that **Logistic Regression achieved the highest accuracy of 98.87%**, followed by Naïve Bayes, Decision Tree, and KNN. The **Random Forest algorithm performed the worst, with an accuracy of 80.85%**. The study highlights that selecting optimal features plays a crucial role in improving model performance.

Limitations

Despite achieving high accuracy, the system has certain limitations. It may suffer from overfitting if too many features are considered. Additionally, the dataset used in the study is limited to 133 columns with predefined symptoms and diseases, which may not cover all possible medical conditions. The study also acknowledges that real-world medical data is often noisy and incomplete, making it challenging to generalize the model effectively.

Future Scope

The authors suggest several improvements for future work, including integrating deep learning techniques for better feature extraction and prediction accuracy. Expanding the dataset with real-time patient data and enhancing model interpretability for healthcare professionals are also recommended. Additionally, they propose incorporating wearable health monitoring systems to collect continuous patient data for more accurate disease prediction.

Title 2: Disease Prediction using Machine Learning

Authors: Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar, and Dr. Shivi Sharma

Publication Date: May, 2021

Proposed System

The proposed system focuses on predicting chronic diseases using machine learning techniques. It aims to provide an automated software solution that can predict diseases based on user-provided symptoms, thereby saving time and money for patients. The system uses a combination of structured and unstructured data from various health-related websites. The framework employs data mining techniques to detect chronic diseases early, and it uses a latent factor model to handle missing data in medical records. The system also consults hospital experts to identify useful features for structured data and uses the random forest algorithm for feature selection in unstructured text files.

Dataset used: Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered. The symptoms of the disease are collected from different health related websites.

Algorithms Used

The primary algorithm used in this study is the **Random Forest Classifier**. The authors also mention the use of data preprocessing techniques such as forward fill for handling null values, data standardization using mean and standard deviation, and splitting the dataset into training and testing sets.

Methodology for Increasing the Accuracy:

To increase accuracy, the authors employed several data preprocessing steps, including checking for null values, converting data into different cases, and standardizing the data. The dataset was split into training and testing sets to ensure the model's robustness. The random forest algorithm was chosen for its ability to handle both structured and unstructured data effectively. The model was trained using chronic disease datasets, and the accuracy was evaluated based on the testing data.

Accuracy Generated:

The accuracy achieved by the random forest classifier for different diseases is as follows:
Diabetes Model: 98.25%, **Breast Cancer Model**: 98.25%, **Heart Disease Model**: 85.25%, **Kidney Disease Model**: 99%, **Liver Disease Model**: 78%

Limitations:

The study does not explicitly mention limitations, but potential limitations could include the reliance on online data sources, which may not always be accurate or comprehensive. Additionally, the model's performance on unstructured text data might be limited by the quality of feature selection. The system may also face challenges in generalizing to different populations or regions due to variations in disease prevalence and healthcare data.

Future Scope:

The authors suggest that future work could focus on improving the accuracy of the model, especially for diseases like liver disease, where the accuracy is relatively lower (78%). Additionally, the system could be expanded to include more diseases and incorporate real-time data from healthcare providers. Further research could also explore the integration of other machine learning algorithms or hybrid models to enhance predictive performance.

Title 3: Human Disease Prediction using Machine Learning Techniques and Real-life Parameters

Authors: K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawanshi

Publication Date: June, 2023

Proposed System:

The study proposes an advanced machine learning-based system for predicting human diseases based on real-life parameters. The system leverages patient data, including symptoms, demographics, and lifestyle factors, to predict diseases accurately. It aims to improve healthcare efficiency by reducing the workload on doctors and enhancing early disease detection. The model integrates structured and unstructured data sources, including medical records and real-time symptom tracking.

Dataset used: https://www.kaggle.com/datasets/itachi9604/diseasesymptomdescription-dataset?select=dataset.csv

Algorithms Used:

The research utilizes multiple machine-learning techniques, including Random Forest, Long Short-Term Memory (LSTM), and Support Vector Machine (SVM). These models are selected based on their ability to classify diseases efficiently and handle large datasets. The Random Forest model, in particular, is highlighted for its superior accuracy in classifying diseases.

Methodology for Increasing Accuracy

The authors employ several optimization techniques to improve model accuracy. These include **hyperparameter tuning in Random Forest**, assigning weighted values to rare symptoms based on geographic distribution, and utilizing LSTM for time-series analysis of patient history. The dataset used for training is sourced from Kaggle and preprocessed to ensure data consistency. The study also applies feature selection techniques to remove irrelevant or redundant data points

Accuracy Generated

The results indicate that **the Random Forest model achieved the highest accuracy at 97%**, outperforming other models such as Weighted KNN (93.5%), Naïve Bayes (94.8%), and SVM (90%). The study highlights the effectiveness of the proposed model in disease classification compared to traditional approaches.

Limitations

The model depends on the availability of high-quality, structured datasets, which may not always be accessible. Additionally, the Random Forest model, while highly accurate, has a higher computational cost compared to simpler algorithms. The research also acknowledges that the model's accuracy could be affected by real-world factors such as missing or inaccurate patient data

Future Scope

The study suggests expanding the dataset to include real-time electronic health records (EHRs) and integrating deep learning models for more complex disease prediction. Future work includes incorporating additional health parameters, such as wearable device data, and refining the model for deployment in hospital management systems.

Title 4: Machine Learning-Based Disease Prediction

Authors: Rakibul Islam, Azrin Sultana and Mohammad Rashedul Islam

Publication Date: July, 2024

Proposed System

The research focuses on predicting chronic diseases using machine learning techniques. The proposed system utilizes various machine learning models to analyze medical data and identify early risk factors associated with diseases such as liver disease, diabetes, cancer, and heart disease. The system aims to improve early diagnosis and aid in preventive healthcare. The study also compares different models to determine the best-performing one.

Dataset Used: Provided different datasets for Liver Disease, Cancer Disease, Brain Disease, Heart Disease, Diabetes disease. Publicly available datasets are :
Skin cancer dataset(Cancer) from Kaggle, Hepatitis C dataset (Liver) From UCI repository, Parkinson disease dataset (Brain) from UCI repository, Heart disease dataset(Heart) from Kaggle, Early-stage diabetes risk prediction dataset (Diabetes) from UCI Repository

Algorithms Used

The study incorporates a variety of machine learning algorithms, including **XGBoost, Support Vector Machines (SVM), Stochastic Gradient Descent (SGD) classifier, Random Forest (RF), Logistic Regression (LR), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks**. The models were implemented using TensorFlow and the scikit-learn library.

Methodology for Increasing Accuracy

The researchers employed several feature selection techniques, including Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR), to improve model accuracy. They also used Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance and ensure better generalization. The models were trained using a 70-30 train-test split, and hyperparameter tuning was performed using grid search.

Accuracy Generated

Among the models tested, XGBoost demonstrated the highest accuracy, with an Area Under the Curve (AUC) score of 91.4% and a precision of 83.1%. Other models such as Random Forest and SVM also performed well, but XGBoost outperformed them in terms of specificity, sensitivity, and F1-score.

Limitations

The research highlights certain limitations, including the reliance on publicly available datasets that may not be fully representative of real-world patient data. Additionally, the study mainly focuses on predictive modeling and does not explore clinical decision support system (CDSS) integration, which could enhance its real-world applicability. The models' generalization ability is also a concern due to dataset constraints.

Future Scope

The authors suggest further research in deep learning techniques, particularly in integrating electronic health records for real-time disease monitoring. They also propose improving the interpretability of models by using explainable AI techniques such as SHAP (Shapley Additive Explanations). The study recommends developing a more comprehensive CDSS that can integrate with hospital management systems for better patient care.

Title: Disease Prediction from Various Symptoms Using Machine Learning.

By: inkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, and Ninad Mehendale

Publication Date: July, 2024

Proposed System: The study introduces a system that predicts diseases based on symptoms, age, and gender. It uses machine learning models to analyze patient data and provide a probable diagnosis, aiming to assist in early detection and timely treatment. The dataset consists of more than 230 diseases, making the model versatile in identifying various illnesses.

Dataset used: The dataset consisting of gender, symptoms, and age of an individual. No Link or source provided

Algorithms Used: Multiple machine learning models were tested, including Decision Trees (Fine, Medium, and Coarse), Gaussian Naïve Bayes, Kernel Naïve Bayes, K-Nearest Neighbors (Fine, Medium, Coarse, and Weighted), Subspace KNN, and RUSBoosted Trees. Among these, Weighted KNN performed the best.

Methodology for Increasing Accuracy: To improve accuracy, the dataset was preprocessed by categorizing symptoms, age, and gender. The data was then split into training and testing sets before being fed into different ML models. Weighted KNN was particularly effective as it assigned greater importance to closer data points, enhancing precision.

Accuracy Achieved: The **Weighted KNN model achieved the highest accuracy of 93.5%**, followed by **Fine KNN (80.3%) and Subspace KNN (73.2%)**. Other models had lower performance, with **RUSBoosted Trees being the least accurate at just 0.5%**.

Limitations: Some models failed to achieve high accuracy due to parameter dependencies. The dataset may not fully account for complex medical cases, rare diseases, or environmental and genetic factors. Additionally, the model relies on symptom-based inputs, which might not always be sufficient for accurate diagnosis.

Future Scope: The system can be enhanced by expanding the dataset, incorporating deep learning models like CNN or LSTM for better pattern recognition, and integrating additional factors such as patient history and external influences. Further development into a user-friendly application could make it a valuable tool in healthcare.