Title: THE PREDICTION OF DEASES USING MACHINE LEARNING.

Link: https://www.researchgate.net/publication/357449131_THE_PREDICTION_OF_DISEASE_USING_MACHINE_LEARNING

Name of Author: Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya.

Proposed System

The proposed system predicts various chronic diseases based on user-inputted symptoms. It processes structured data through machine learning models, utilizing Naïve Bayes for disease prediction, KNN for classification, logistic regression for feature extraction, and decision trees for data segmentation. The system provides accurate disease predictions, enhancing healthcare decision-making.

Algorithm Used

The system employs **Naïve Bayes**, **KNN (K-Nearest Neighbors)**, **Logistic Regression**, and **Decision Tree** algorithms. Naïve Bayes is used for disease probability estimation, KNN for classification based on symptom similarity, logistic regression for key feature extraction, and decision trees for organizing and refining datasets.

Methodology Used to Increase Accuracy

To improve accuracy, the system evaluates model performance using **precision, recall, accuracy, and F1-score** metrics. Feature selection techniques enhance prediction reliability, while different algorithms like Decision Tree and Naïve Bayes optimize classification accuracy.

Limitations

The system relies on structured input data, limiting flexibility in handling unstructured medical records. It may not generalize well to rare diseases due to data constraints. Additionally, incorrect or incomplete symptom inputs can affect prediction reliability.

Future Scope

Future improvements include integrating deep learning for better pattern recognition, incorporating unstructured data like medical reports, and expanding the disease database. Enhancements in real-time healthcare monitoring and IoT-based symptom tracking are potential advancements.

Conclusion

The disease prediction system effectively identifies diseases based on symptoms, offering a **user-friendly web-based platform** accessible from anywhere. With a high accuracy rate, the system aids healthcare professionals in early diagnosis and personalized treatment recommendations, ultimately improving patient care.

Title: Comparing Different Supervised Machine Learning Algorithms for Disease Prediction.

Proposed System:

This research compares supervised machine learning algorithms for disease prediction using medical data. It analyzes algorithms such as SVM, Naïve Bayes, Random Forest, Decision Tree, Logistic Regression, KNN, and ANN to determine their effectiveness. The study reviews 48 research papers, ensuring a standardized comparison across different diseases. The findings help researchers identify the most suitable algorithms based on accuracy, sensitivity, and specificity.

Algorithms Used:

The study evaluates Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN).

Methodology for Increasing Accuracy:

To improve accuracy, multiple algorithms are tested on the same dataset to avoid selection bias. Cross-validation methods, such as 5-fold and 10-fold validation, enhance generalization. The study also examines algorithm strengths and weaknesses in different medical applications.

Accuracy Generated:

Random Forest (RF) achieved the highest accuracy in 53% of studies, followed by Support Vector Machine (SVM) at 41%. Different diseases showed varying algorithm performance—ANN was most effective for breast cancer, while SVM performed best for diabetes and heart disease.

Limitations:

Dataset variability across studies affects algorithm performance. The study does not consider hyperparameter tuning, which impacts accuracy. A lack of standardized feature selection introduces inconsistencies in algorithm comparison.

Future Scope:

Future research should explore deep learning models like CNNs and RNNs for improved accuracy. Standardizing datasets will enhance comparability. Developing hybrid models combining multiple algorithms can further improve disease prediction.

Marwadi University
Marwadi Chandarana Group

Disease Prediction by Machine Learning Over Big Data from Healthcare Communities.

Publication Date and Dataset Used.

The paper, titled *Disease Prediction by Machine Learning Over Big Data From Healthcare Communities*, was written in [insert date from paper]. It explores how machine learning (ML) techniques can improve disease prediction using big data from healthcare communities. The study discusses various ML algorithms such as decision trees, random forests, and neural networks, emphasizing their accuracy in predicting diseases based on patient data. It highlights the importance of feature selection, data preprocessing, and handling imbalanced datasets to enhance predictive performance. The paper also discusses challenges such as data privacy, ethical concerns, and computational efficiency in implementing ML models in real-world healthcare settings.

This study can be accessed at [insert location or DOI if available], and it utilizes a dataset comprising patient medical records and diagnostic reports from healthcare communities, focusing on structured and unstructured data integration for better prediction accuracy.

Proposed System:
This study presents a machine learning approach for predicting chronic disease outbreaks using big data from healthcare communities. The proposed system integrates structured and unstructured hospital data to improve accuracy. It introduces a CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm, which processes electronic health records (EHR), medical images, and clinical notes. The study focuses on cerebral infarction, a high-risk chronic disease, using real-life hospital data collected in central China between 2013 and 2015. The system also employs a latent factor model to reconstruct missing data, ensuring completeness and reliability.

Algorithms Used:
The research utilizes Naïve Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree (DT), and deep learning models such as CNN-UDRP (unimodal) and CNN-MDRP (multimodal).

Methodology for Increasing Accuracy:
The study integrates structured patient data (demographics, lab results) with unstructured text data (medical notes, diagnoses) to enhance prediction accuracy. CNN extracts features from text, while structured data undergoes preprocessing and statistical analysis. A latent factor model fills in missing values, and cross-validation ensures model robustness.

Accuracy Generated:

The CNN-MDRP model achieved an accuracy of 94.8%, surpassing other traditional machine learning models. It also exhibited a faster convergence speed compared to the unimodal CNN-UDRP model.

Limitations:

The system relies on historical hospital data, which may not generalize well to new cases. Variability in regional disease patterns affects model performance. The study does not explore hybrid models beyond CNN-based approaches.

Future Scope:

Future work should integrate more deep learning techniques, such as transformers, for better feature extraction. Expanding dataset diversity across multiple regions can improve generalization. Combining different AI models, including reinforcement learning, may further enhance disease prediction accuracy.

Heart Disease Prediction Using Machine Learning Algorithms

Proposed System:
This research focuses on developing a heart disease prediction system (HDPS) using machine learning techniques. The system analyzes patient medical records to predict the likelihood of heart disease based on factors such as age, chest pain, blood pressure, and fasting blood sugar levels. The study uses a dataset from the UCI repository containing 304 patient records. The model aims to enhance early diagnosis, reduce medical costs, and improve patient outcomes by leveraging multiple machine learning algorithms. The study highlights the effectiveness of using multiple classifiers rather than a single algorithm to achieve better accuracy.

Algorithms Used:
The research employs Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier for heart disease prediction.

Methodology for Increasing Accuracy:
Data preprocessing is performed to clean and normalize the dataset. The study extracts significant features, splits the data into training and testing sets, and applies multiple classifiers. The combination of three different machine learning techniques improves accuracy by reducing bias. The study also compares its results with previous works to validate performance improvements.

Accuracy Generated:
Among the models tested, K-Nearest Neighbors (KNN) achieved the highest accuracy of 88.52%, followed by Logistic Regression. The overall model accuracy averaged 87.5%, outperforming previous studies that achieved around 85%

.

Limitations:
The dataset used is relatively small, limiting generalization to larger populations. The study does not explore deep learning techniques, which could further enhance performance. The system relies on predefined medical attributes, which may not capture all relevant risk factors.

Future Scope:
Future research should incorporate deep learning techniques such as neural networks for improved feature extraction. Expanding the dataset with more diverse patient records could improve generalization. Integrating real-time health monitoring data from wearable devices may further enhance prediction accuracy.

Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach

Proposed System:
This research develops a machine learning-based system to predict hospitalizations due to chronic diseases, specifically heart disease and diabetes, using Electronic Health Records (EHR). The system formulates the prediction task as a binary classification problem and explores multiple machine learning methods to balance accuracy with interpretability. It introduces two novel methods: K-LRT (a likelihood ratio test-based model) and Joint Clustering and Classification (JCC), which clusters patients based on hidden patterns and adapts classifiers to each cluster. The models are trained and validated on large-scale datasets from the Boston Medical Center.

Algorithms Used:
The study employs kernelized and sparse Support Vector Machines (SVM), sparse Logistic Regression, Random Forests, K-LRT, and JCC.

Methodology for Increasing Accuracy:
The system integrates structured and unstructured medical data, utilizing feature selection techniques to improve classification performance. Sparse classifiers are used to enhance model interpretability while retaining predictive power. The study also applies clustering methods to identify patient subgroups with shared characteristics, optimizing the classification process for each subgroup.

Accuracy Generated:
The Random Forest model achieved the highest accuracy, with an area under the ROC curve (AUC) of 81.6% for heart disease and 84.5% for diabetes. The K-LRT and JCC models provided interpretable results while maintaining competitive accuracy.

Limitations:
The study relies on historical hospital data, which may not generalize well to unseen cases. The dataset is highly imbalanced, with fewer hospitalized cases, which could impact model performance. Some models prioritize interpretability over pure predictive accuracy.

Future Scope:

Future research should incorporate real-time health monitoring data to improve early disease detection. Deep learning models, such as transformers and recurrent neural networks, could be explored for improved accuracy. Expanding datasets across diverse populations may enhance generalizability.

Chronic Kidney Disease Prediction Using Machine Learning Methods

Publication Date and Dataset Used:

The paper *"Chronic Kidney Disease Prediction Using Machine Learning"*, authored in March 2023, explores how machine learning algorithms—such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—can be effectively applied to predict chronic kidney disease (CKD). It highlights the importance of early detection in improving patient outcomes and demonstrates that the Random Forest classifier achieved the highest accuracy of 99.2%. The study emphasizes data preprocessing, feature selection, and performance metrics (accuracy, precision, recall) as critical components. The research is available through Jain University and utilizes a publicly accessible CKD dataset from the UCI Machine Learning Repository. The work is accessible via internal project documentation or institutional repositories from Jain University.

Proposed System:
This study presents a machine learning-based approach for predicting Chronic Kidney Disease (CKD) using clinical data. The system aims to enhance early diagnosis by addressing challenges in data preprocessing, missing value handling, and feature selection. The research evaluates 11 machine learning models to identify the most effective classifier for CKD prediction. It also emphasizes the importance of domain knowledge in selecting relevant features and reducing bias in model training.

Algorithms Used:
The study utilizes Decision Tree, Random Forest, XGBoost, Extra Trees, AdaBoost, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, Gaussian Naïve Bayes, and a classical Neural Network.

Methodology for Increasing Accuracy:
To improve model performance, missing values are handled using the KNN-imputer method instead of constant substitution. Feature selection is based on statistical analysis and domain expertise to retain only the most relevant attributes. Hyperparameter tuning is performed using grid search and genetic algorithms to optimize classifiers.

Accuracy Generated:
Random Forest and Extra Trees classifiers achieved 100% accuracy, outperforming other models. XGBoost also delivered strong results, with cross-validation confirming the robustness of the top-performing models.

Limitations:
The dataset is relatively small, which may limit generalization to broader populations. While the study achieves high accuracy, real-world implementation requires validation on larger and more diverse datasets. The reliance on static clinical data excludes real-time patient monitoring.

Future Scope:
Future research should incorporate real-time data from wearable devices to improve prediction capabilities. Expanding datasets to include genetic and lifestyle factors can enhance model generalization. Exploring deep learning methods, such as transformers, could further refine CKD risk prediction.

Study Title	Algorithms Used	Dataset Used	Key Findings	Limitations	Future Scope
The Prediction of Diseases Using Machine Learning	Naïve Bayes, KNN, Logistic Regression, Decision Tree	Structured patient data from healthcare sources	Machine learning models improve chronic disease prediction accuracy	Limited ability to handle unstructured data and rare diseases	Integration of deep learning for better pattern recognition
Comparing Different Supervised Machine Learning Algorithms for Disease Prediction	SVM, Naïve Bayes, Random Forest, Decision Tree, Logistic Regression, KNN, ANN	Medical datasets from 48 research papers	Random Forest achieved the highest accuracy (53% of studies), ANN performed best for breast cancer, SVM for diabetes and heart disease	Dataset variability affects algorithm performance, lacks hyperparameter tuning	Deep learning models like CNNs and RNNs could improve accuracy
Disease Prediction by Machine Learning Over Big Data from Healthcare Communities	CNN, Naïve Bayes, KNN, Decision Tree	Hospital data from central China (2013-2015)	CNN-MDRP achieved 94.8% accuracy, outperforming other models	Regional disease pattern variations affect model performance	Expansion of dataset diversity and integration of hybrid AI models
Heart Disease Prediction Using Machine Learning Algorithms	Logistic Regression, KNN, Random Forest	UCI repository (304 patient records)	KNN achieved highest accuracy (88.52%), model improves early diagnosis	Small dataset limits generalization, lacks deep learning techniques	Incorporation of deep learning and real-time health monitoring
Predicting Chronic Disease Hospitalizations from Electronic Health Records	SVM, Logistic Regression, Random Forest, K-LRT, JCC	Electronic Health Records from Boston Medical Center	Random Forest achieved 81.6% AUC for heart disease, 84.5% for diabetes	Imbalanced dataset affects model performance	Real-time monitoring and deep learning models for enhanced prediction
Chronic Kidney Disease Prediction Using Machine Learning	Decision Tree, Random Forest, XGBoost, Extra Trees, AdaBoost, KNN, SVC, Logistic Regression, Naïve Bayes, Neural Network	UCI CKD dataset	Random Forest and Extra Trees achieved 100% accuracy	Small dataset may not generalize well	Inclusion of real-time data from wearable devices and genetic factors