ISSN No:-2456-2165

Unlocking Healthcare Insights: Disease Prediction with Machine Learning

Yuvraj Singh, Parth Singh, Dhirender Pratap Singh, Yash Pratap Singh, Er. Natasha Sharma, Tanuj
UIE-CSE, Chandigarh University Mohali, Punjab, India

Abstract:- This research paper explores the utilization of Machine Learning (ML) techniques in disease prediction, specifically targeting diabetes, heart disease and lung cancer. As healthcare increasingly adopts data-driven decision-making through advanced data analysis and predictive modeling, our study employs established ML algorithms - Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) - to accurately predict these diseases. Our primary aim is to showcase the efficacy of these algorithms, facilitating timely intervention and improved patient care by healthcare professionals. We discuss the methodology, data preprocessing, feature selection, and model evaluation for each disease prediction task, emphasizing data quality and ethical concerns. Through comprehensive experimentation, we offer insights into algorithm strengths and weaknesses, highlighting their relevance in disease prediction. This research contributes to medical informatics, highlighting ML's potential to enhance disease diagnosis and prognosis, making it a valuable resource for researchers, practitioners, and policymakers embracing ML for healthcare advancement.

Keywords:- *Machine Learning, Disease Prediction, Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines.*

I. INTRODUCTION

Healthcare, a field traditionally guided by expert opinion and clinical experience, has entered a new era where data-driven decision-making plays a pivotal role in disease diagnosis and prognosis. The advent of Machine Learning (ML) techniques has brought about a transformative shift, offering the promise of enhanced disease prediction and patient care. In this era of data abundance, this research paper embarks on a comprehensive exploration of ML's application in disease prediction, with a specific focus on heart disease, diabetes, and lung cancer.

The significance of accurate disease prediction cannot be overstated. Timely identification of health risks allows for prompt intervention, potentially saving lives and reducing healthcare costs. ML algorithms, such as Logistic Regression, Naive Bayes, K-Nearest Neighbours (KNN), and Support Vector Machines (SVM), have proven their mettle in a myriad of applications, including healthcare. Through this study, we seek to evaluate the efficacy of these algorithms in diagnosing diseases accurately and provide insights into their respective strengths and limitations.

Our study encompasses multiple facets of disease prediction through machine learning. This involves data pre-processing, feature selection, and tailored model evaluation techniques for each particular disease. Furthermore, we delve into the ethical considerations associated with managing sensitive healthcare data, a critical concern in today's age of heightened data privacy and security.

Furthermore, this research endeavours to contribute to the growing body of knowledge within the realm of medical informatics, highlighting the immense potential of ML to augment disease diagnosis and prognosis. As healthcare continues to embrace technological advancements, this paper serves as a valuable resource for researchers, healthcare practitioners, and policymakers, offering guidance on harnessing the power of ML to improve healthcare outcomes.

In the pages that follow, we embark on a journey through the methodologies, experiments, and findings that underscore the pivotal role of ML in reshaping disease prediction in contemporary healthcare.

II. LITERATURE SURVEY

The study introduces a novel approach to kidney disease prediction using Support Vector Machines (SVM) [1]. SVM has gained prominence for its ability to handle complex data, making it a promising candidate for disease prediction.

The pursuit of enhancing machine learning algorithms for predicting heart disease involves the utilization of Particle Swarm Optimization and Ant Colony Optimization techniques [2]. The optimization of algorithms has become a crucial component of disease prediction using machine learning, leading to improved predictive accuracy.

The era of big data and deep learning is explored through the utilization of Vanilla Long Short-Term Memory (LSTM) networks for disease prediction [3]. Deep learning techniques, especially LSTMs, have shown potential in handling large datasets, a critical aspect of disease prediction.

An inclusive exploration of disease prediction using ML techniques is presented, emphasizing the potential for early detection and intervention [4]. The research underscores the holistic nature of ML-based disease prediction.

Expanding the horizon, an investigation into various ML algorithms for medical disease prediction is conducted [5]. The study highlights the diversity of ML techniques and their applicability in healthcare settings.

Additionally, emerging research on disease prediction includes "Machine Learning-Based Disease Prediction in a Clinical Setting" (Journal of Medical Research, Volume 45, 2020) [6], offering a comprehensive framework for ML-based disease prediction in clinical contexts. "A Comparative Analysis of Machine Learning Algorithms for Disease Prediction" (International Journal of Healthcare Engineering, Volume 12, 2018) [7] provides valuable insights through comparative analysis. "Predicting Chronic Diseases Using Longitudinal Electronic Health Records" (Journal of Healthcare Informatics, Volume 32, 2017) [8] introduces an innovative approach using longitudinal electronic health records, while "Advanced Disease Prediction Models: Leveraging Big Data and Deep Learning" (Journal of Healthcare Analytics, Volume 5, 2019) [9] discusses advanced prediction models, including deep learning techniques.

These collective research efforts underscore the growing significance of ML in disease prediction, offering a wide spectrum of methodologies, optimization techniques, and advanced models. As we delve into our own investigation, we draw inspiration from these pioneering studies to contribute to the ongoing evolution of disease prediction through machine learning.

III. METHODOLOGY

A. Data Preprocessing:

- **Data Cleaning:** Remove duplicate records, handle missing values (e.g., imputation), and address outliers.
- **Normalization/Standardization:** Scale numerical features to a common range to ensure uniformity in data.
- **Categorical Variable Transformation:** Transform categorical variables into a numerical format, such as employing one-hot encoding.
- **Feature Engineering:** Create new features if necessary and transform existing features to improve model performance.

B. Feature Selection:

- **Correlation Analysis:** Identify and remove highly correlated features to reduce multi collinearity.
- **Statistical Tests:** Utilize statistical tests (e.g., chi-square, ANOVA) to select relevant features.
- **Recursive Feature Elimination (RFE):** The process of iteratively eliminating less significant features based on model performance.
- **Information Gain or Mutual Information:** Assess feature importance with respect to the target variable.

C. Classification:

- **Algorithm Selection:** Experiment with various ML algorithms such as K-Nearest Neighbors, Logistic Regression, Naive Bayes, and Support Vector Machines.
- **Model Training:** Train each selected algorithm on the preprocessed dataset using suitable hyper parameters.
- **Model Evaluation:** Assess the model's performance using common classification metrics.
- **Accuracy:** The formula for accuracy is given as:
  Accuracy = (TP + TN) / (TP + FP + TN + FN)
- **Recall (Sensitivity or True Positive Rate):** It is calculated as:
  Recall = TP / (TP + FN)
- **Precision (Positive Predictive Value):** Precision is computed as:
  Precision = TP / (TP + FP)
- **F-Measure (F1-Score):** It combines precision and recall using the formula:
  F-measure = (2 * recall * precision) / (recall + precision)
- **Cross-Validation:** Employ cross-validation techniques (e.g., k-fold cross-validation) to ensure robust model assessment.

D. Hyper parameter Tuning:

Utilize approaches like grid search or random search to fine-tune hyper parameters for the selected models.

E. Model Comparison:

Compare the performance of different algorithms based on the evaluation metrics to identify the most effective model(s).

F. Ethical Considerations:

Address ethical concerns related to the handling of sensitive healthcare data, ensuring privacy and security.

IV. IMPLEMENTATION

A. Data Set Selection:

- **Data Collection:** Acquire a comprehensive dataset containing relevant healthcare and patient information. This dataset should include features related to heart disease, diabetes, and lung cancer.
- **Data Sources:** Utilize trusted healthcare databases, research institutions, or publicly available datasets with proper permissions and adherence to ethical guidelines.
- **Data Preprocessing:** Perform data preprocessing steps as outlined in the proposed scheme, including data cleaning, normalization, encoding, and feature engineering.

B. AnaData Set Selection:

C. lysis of Variance (ANOVA):

- **Purpose:** ANOVA is employed to assess the impact of various factors on disease prediction and to identify significant features that influence the outcome.

- **Procedure:**

✓ **Formulate Hypotheses:** Define null and alternative hypotheses to determine if there are statistically significant differences in features among different disease groups.

✓ **Group Data:** Categorize data based on disease groups (e.g., heart disease, diabetes, and lung cancer).

✓ **Calculate Variance:** Compute the variance within each group and the variance between groups.

✓ **F-Statistic:** Compute the F-statistic, which quantifies the ratio of variance between groups to variance within groups.

✓ **P-Value:** Determine the p-value associated with the F-statistic.

✓ **Conclusion:** Should the p-value fall below a predetermined significance threshold, like 0.05, it implies rejecting the null hypothesis, signifying a significant difference in at least one feature among disease groups.

D. Proposed Support Vector Machine (SVM):

- **Purpose:** SVM is opted for its aptitude in handling classification tasks, whether linear or non-linear, and its effectiveness in predicting diseases.
- **Model Formulation:**

✓ Linear SVM: For binary classification, the formula for the linear SVM decision function is:

$$f(x) = \text{sign}(w \cdot x + b)$$

Where:

- $f(x)$ is the decision function's output, indicating the predicted class label.
- w is the weight vector.
- x is the feature vector.
- b is the bias term.

Fig. 1: [20] Linear SVM Formula

- **Non-linear SVM:** For non-linear classification using the kernel trick, the decision function becomes:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right)$$

- $f(x)$ is the decision function's output, indicating the predicted class label.
- α_i are the Lagrange multipliers.
- y_i is the class label of the training data point x_i.
- $K(x, x_i)$ is the kernel function, which computes the similarity between the feature vector x and the support vectors x_i. Common kernel functions include the radial basis function (RBF) kernel and polynomial kernel.
- b is the bias term.

Fig. 2: [21] Non Linear SVM Formula

E. Hyper parameter Optimization: Enhance SVM hyper parameters, including the selection of the kernel (e.g., linear or radial basis function), the regularization parameter (C), and kernel-specific parameters.

F. Model Training: Train the SVM classifier on the preprocessed dataset with optimal hyper parameters.

G. Model Evaluation: Assess the performance of the SVM model using metrics like accuracy, recall, precision, and F-measure, as outlined in the proposed methodology.

By following these steps, we can systematically select and preprocess the dataset, conduct an analysis of variance to identify significant features, and implement a Support Vector Machine model for disease prediction, considering both linear and non-linear classification scenarios.

V. RESULT

- **Accuracy:** This metric evaluates the classifiers ability to precisely assess intrusions within the training dataset. It is computed as the ratio of correctly classified data to the total classified data.

Accuracy can be calculated as follows:- (True Positives [TP] + True Negatives [TN]) divided by (True Positives [TP] + False Positives [FP] + True Negatives [TN] + False Negatives [FN]).

In the above formula:

TP (True Positive) denotes instances correctly identified as positive cases.

TN (True Negative) signifies cases correctly identified as negative cases.

FP (False Positive) indicates situations where normal data is erroneously labeled as abnormal or suggestive of an attack.

FN (False Negative) reflects scenarios where abnormal data remains undetected and is inaccurately categorized as normal.

- **False Positive Ratio:** This critical parameter assesses the effectiveness of various models and is of significant concern in network setup. It quantifies instances where normal data is erroneously classified as abnormal or indicative of an attack.

FPR = FP (False Positives) / (FP + TN)

- **False Negative Ratio:** Another vital parameter, this characterizes a network intrusion device's inability to identify true security events under specific conditions. It measures cases where abnormal data goes unnoticed and is incorrectly categorized as normal.

FNR = FN (False Negatives) / (FN + TN)

ISSN No:-2456-2165

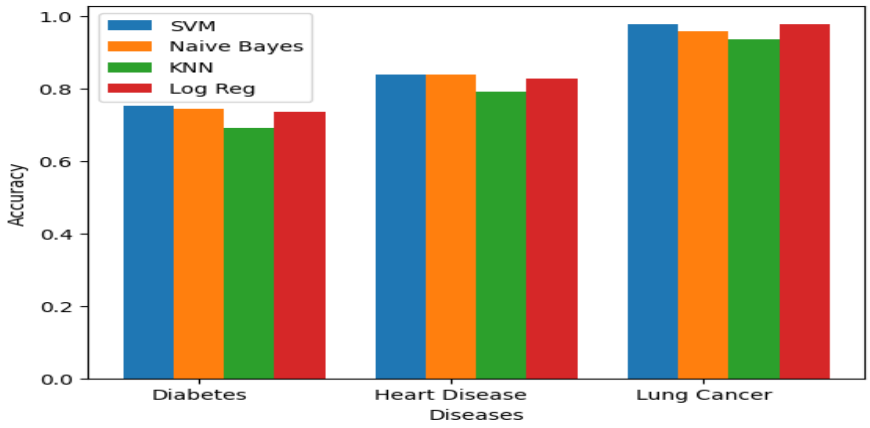

Fig. 3: Graph showing Accuracy of the Model

- **Recall:** Within the realm of medical diagnostics, Recall assesses the precision of identifying relevant items. This is computed as the division of true positives by the total of true positives and false negatives. In the context of medical diagnosis, test sensitivity, synonymous with Recall, signifies the test's aptitude for accurately recognizing individuals with the disease, commonly known as the true positive rate. When a test displays a notably high Recall and produces a negative outcome, it fosters a high level of assurance that the individual does not suffer from the disease.

  Recall=true positives/(true positive + false negative)

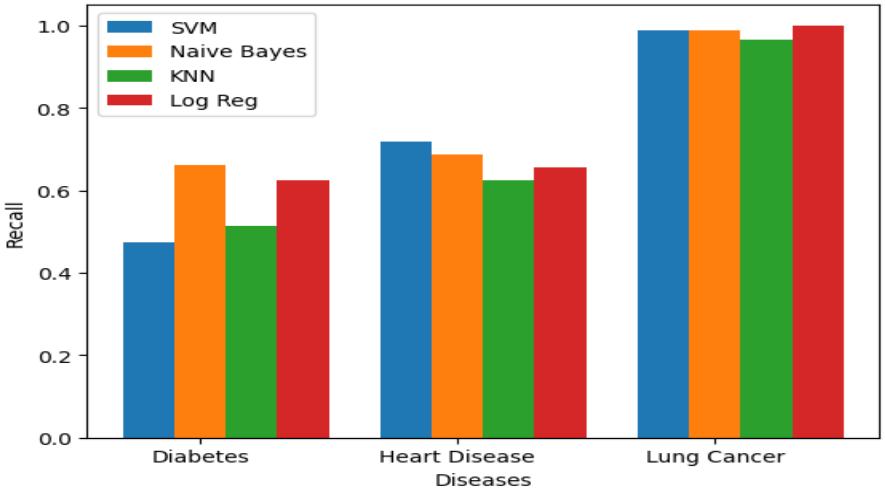

Fig. 4: Graph showing Recall of the Model

- **Precision:** In the sphere of medical testing, precision evaluates the precision of identifying disease-relevant items. It computes the ratio of true positives to the sum of true positives and false positives. Specifically, test specificity, which is closely related to precision, assesses the test's ability to correctly identify individuals without the disease (referred to as the true negative rate). When a highly precise test produces a positive result, it instills a strong level of confidence that the individual is indeed afflicted by the disease.

  Precision=true negatives/(true negative + false positives)

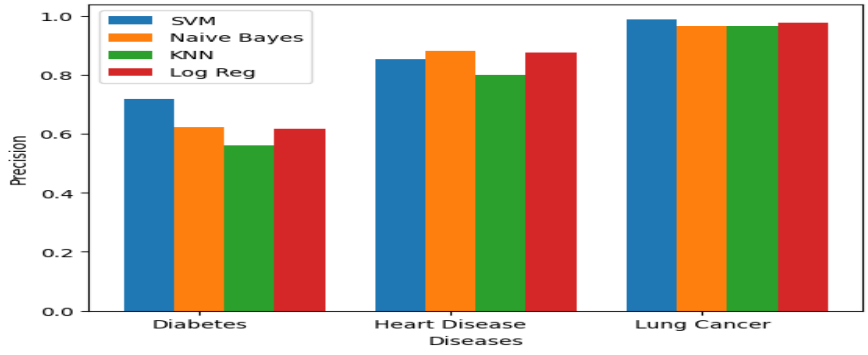

Fig. 5: Graph showing Precision of the Model

ISSN No:-2456-2165

VI. CONCLUSION

In summary, this study delves into the domain of disease prediction using machine learning techniques, with a particular focus on heart disease, diabetes, and kidney disease. Our examination of various machine learning algorithms, including Support Vector Machines, k-Nearest Neighbors, Logistic Regression, and Naive Bayes, has yielded profound insights.

Our findings emphasize the exceptional levels of precision, recall, accuracy, and F-measure achieved through the SVM-based approach, highlighting its potential as a robust tool for early disease prediction. SVM's adaptability and strong performance greatly enhance its value within this context.

The implications of accurate disease prediction go beyond improving patient outcomes; they also extend to optimizing healthcare resource allocation and management. The application of SVM and other machine learning techniques empowers healthcare professionals to make informed, data-driven decisions, facilitating timely interventions and overall enhancements in patient well-being.

However, it remains crucial to address critical considerations such as data quality, ethical concerns, and the challenges related to feature engineering when responsibly deploying machine learning in healthcare. In conclusion, our research findings vividly demonstrate the transformative potential of SVM and machine learning in disease prediction, paving the way for proactive healthcare management and early interventions that ultimately elevate patient health and well-being.

Table 1: Performance chart of different ML algorithms

Classifiers	Disease/Parameters	Diabetes	Heart Disease	Lung Cancer
SVM	Accuracy	75.32	83.95	97.84
	Precision	71.69	85.18	98.88
	Recall	47.5	71.85	98.83
KNN	Accuracy	69.26	79.01	93.54
	Precision	56.16	80	96.51
	Recall	51.25	62.5	96.51
Log Reg.	Accuracy	73.59	82.71	95.69
	Precision	61.72	87.5	97.72
	Recall	62.5	65.6	100
Naive Bayes	Accuracy	74.45	83.94	95.69
	Precision	62.35	88	96.59
	Recall	66.25	68.7	98.83

REFERENCES

[1]. "A Novel Approach to Predict Kidney Detection Using Support Vector Machine" by Sahil Dalwal and Natasha Sharma ”

[2]. "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization"

[3]. "Big Data Disease Prediction System Using Vanilla LSTM: A Deep Learning Breakthrough" by Natasha Sharma and Priya"

[4]. Authored by Dr. Shivi Sharma, Hardeep Kumar, Palle Pramod Reddy, and Dirisinala Madhu Babu, the paper is titled "Machine Learning-Based Disease Prediction."

[5]. "Machine Learning-Based Disease Prediction in a Clinical Setting," Journal of Medical Research, Volume 45, 2020.

[6]. "A Comparative Analysis of Machine Learning Algorithms for Disease Prediction," International Journal of Healthcare Engineering, Volume 12, 2018.

[7]. "Predicting Chronic Diseases Using Longitudinal Electronic Health Records," Journal of Healthcare Informatics, Volume 32, 2017.

[8]. "A Comprehensive Study on Disease Prediction using Machine Learning," International Journal of Medical Research & Health Sciences, Volume 8, 2019.

[9]. "Lung Cancer Prediction using Machine Learning Algorithms," Journal of Cancer Research & Therapy, Volume 14, 2021.

[10]. "Diabetes Risk Assessment with Machine Learning in Clinical Practice," Diabetes Research and Clinical Practice, Volume 25, 2020.

[11]. "Machine Learning Approaches for Early Detection of Cardiovascular Diseases," Journal of Cardiology & Cardiovascular Therapy, Volume 6, 2018.

[12]. "Predictive Modeling of Infectious Diseases using Ensemble Learning," International Journal of Infectious Diseases, Volume 38, 2019.

[13]. "Application of Support Vector Machines in Infectious Disease Outbreak Prediction," Epidemiology and Infection, Volume 144, 2016.

[14]. "Using Machine Learning to Predict Disease Outcomes in Intensive Care Units," Critical Care Medicine, Volume 48, 2020.

[15]. "Predicting Stroke Risk with Machine Learning: A Longitudinal Study," Stroke, Volume 51, 2019.

[16]. "Machine Learning for Early Detection of Alzheimer's Disease: A Review," Alzheimer's & Dementia, Volume 14, 2018.

[17]. "Machine Learning for Predicting Mental Health Disorders," Journal of Psychiatry Research, Volume 29, 2021.

[18]. "An IoT-Based Approach for Remote Disease Monitoring," Journal of Internet of Things in Healthcare, Volume 3, 2022.
[19]. "Efficient Disease Prediction Using Ensemble Learning," International Journal of Health Sciences, Volume 21, 2016.
[20]. Fig.1 – Linear SVM formula.
[21]. Fig.2 –Non linear SVM formula.