COMPARE AND CONTRAST DIFFERENT SUPERVISED MACHINE LEARNING FOR DISEASE PROGNOSIS AND PREDICTION

A thesis submitted to the graduate faculty in partial fulfillment of the requirement of the degree of

BACHELOR OF SCIENCE

By:

Akash Santra (358),
Balraj N(333),
Subhadip Bardhan(201).
Honours: B.Sc Computer Science.
(2020-2023)

Supervised By:

Prof Sarbajit Manna
Assistant Professor,
Dept. of Computer Science(RKMV)

Ramakrishna Mission Vidyamandira

Belur Math,Howrah

CERTIFICATE

This is to certify that we have completed the B. Sc Project Report titled "**Compare and contrast of different supervised machine learning for disease prognosis and prediction**" as a partial fulfillment of the Bachelor of Computer Science degree requirements at the Department of Computer Science of Ramakrishna Mission Vidyamandira, Belur Math. The project was carried out during our 6th semester from January 2023 to May 203 under the guidance of Prof. Sarbajjit Manna, Assistant Professor at the Department of Computer Science, Ramakrishna Mission Vidyamandira, Belur Math. The work presented in this report is an authentic record of our own efforts.

The matter presented in this Project Report has not been submitted for the award of any other degree elsewhere.

Signature of the Students:-

Akash Santra Balraj N Subhadip Bardhan

This is to certify that the above statement made by the students is true to the best of my knowledge.

- Signature of the Supervisor

Date:

- Signature of the Head of Dept

Date:

- Official Address with Seal: Signature of the External Examiner

Acknowledgement

Expressing gratitude and showing appreciation towards individuals who have contributed to our personal or professional growth is a fundamental part of human interaction. In this statement, we express our gratitude towards several individuals who have played a crucial role in our successful completion of a challenging project.

Firstly, we would like to extend our heartfelt gratitude to Professor Sarbajit Manna, our project guide. Professor Manna's exceptional guidance was instrumental in helping us navigate through the challenging stages of the project. He not only extended help in every possible way but also ensured that we remained on track and successfully completed the project. Professor Manna's dedication and commitment to our project were remarkable, and we are deeply grateful for his unwavering support.

Furthermore, we would also like to express our gratitude to all the professors in our department whose teachings have contributed significantly to our growth and learning. Their valuable insights and knowledge have enabled us to apply our theoretical knowledge to practical applications such as this project. We are indebted to them for their invaluable contributions, which have played a vital role in shaping our academic careers.

We would also like to thank our classmates who have engaged in numerous discussions, enhancing the quality of our work. The exchange of ideas and perspectives during group discussions has provided us with valuable insights and helped us think outside the box. We are grateful to our classmates for their active participation, which has contributed immensely to the success of our project.

Additionally, we would like to express our sincere appreciation to Sri Sanjib Kumar Basu, our laboratory attendant, whose unwavering support and guidance kept us motivated during difficult times. Sri Basu's tireless efforts and dedication to his work have been a source of inspiration for us, and we are grateful for his invaluable contributions to our project.

Last but not least, we would like to acknowledge our Revered Principal, Swami Mahaprajnananda, for providing us with the necessary resources and workspace to complete this project. His support and encouragement have been a crucial factor in our project's success, and we are deeply grateful for his guidance and mentorship.

In conclusion, we would like to express our heartfelt gratitude to Professor Sarbajit Manna, all the professors in our department, our classmates, Sri Sanjib Kumar Basu, and our Revered Principal, Swami Mahaprajnananda, for their invaluable contributions to our project. We are deeply grateful for their guidance, support, and mentorship, which have been instrumental in shaping our academic careers and enabling us to achieve our goals.

TABLE OF CONTENTS

PAGE

Abstract

Machine learning (ML) based healthcare data analysis and recommendation systems are becoming increasingly popular for the early detection and prevention of chronic diseases like diabetes and cervical cancer. **The Pima Indians Diabetes Dataset [1]** is commonly used for diabetes diagnosis and prediction, while **The Cervical Cancer Dataset [2]** is used for cervical cancer analysis and prediction. The first step in this process is data preprocessing, followed by feature selection, and then the dataset is split into training and testing datasets. Metrics namely accuracy, precision, recall, F1-score, and area under the curve (AUC) are used to evaluate the performance of the classification algorithm. By employing various ML algorithms, these frameworks can provide accurate and timely diagnosis, which can significantly improve patient outcomes.

Chapter 1 : INTRODUCTION

1.1 Introductory Discussion

Diabetes and cervical cancer are two prevalent diseases that affect millions of people worldwide. Diabetes is a chronic metabolic disorder that results in high blood sugar levels due to the body's inability to produce enough insulin or use it effectively. Cervical cancer, on the other hand, is a type of cancer that affects the cells of the cervix, which is the lower part of the uterus that connects to the vagina.

Machine learning (ML) algorithms can be used to predict and prognosis these diseases by analyzing large amounts of patient data to identify patterns and trends that may be indicative of the presence or progression of the disease. In the case of diabetes, ML algorithms can analyze data such as blood glucose levels, weight, age, family history, and other medical conditions to predict the likelihood of developing diabetes or to assess the severity of the disease in those already diagnosed.

ML algorithms can also be used to predict the risk of developing cervical cancer by analyzing patient data such as age, sexual history, family history, and other risk factors. Additionally, ML algorithms can help to identify early signs of cervical cancer by analyzing patient data such as Pap test results and other medical information.

Prognosis is another area where ML algorithms can be useful in both diabetes and cervical cancer. ML algorithms can analyze patient data such as disease stage, treatment history, and other medical information to predict the likelihood of disease progression or recurrence. This information can be used by healthcare providers to develop personalized treatment plans and monitor patient progress over time.

In summary, ML algorithms can be valuable tools for predicting and prognosis of diseases like diabetes and cervical cancer by analyzing large amounts of patient data to identify patterns and trends that may be indicative of the presence or progression of the disease. These tools can help healthcare providers to develop personalized treatment plans and monitor patient progress over time.

1.2 Problem Definition

The following problems were identified and extracted from various papers and research in the literature and have been addressed in this project:

- Firstly, we observed that there was a lack of comparison between the different models used in previous studies. This made it difficult to determine which model was the most effective for a particular disease prognosis and prediction.
- Secondly, we noticed a lack of comparison and contrast between various hyperparameter tuning methods used to obtain the best performing ML model.
- Finally, we observed that previous studies had implemented machine learning models directly to the dataset after minimal preprocessing, without taking into consideration the potential impact of preprocessing on model performance.

1.3 Contribution

The contributions of this project include:

- Compare and contrast different supervised machine learning models for disease prognosis and prediction, which has the potential to improve the accuracy of disease diagnosis and treatment outcomes.
- A comparison of various machine learning models commonly used in disease prognosis and prediction studies, enabling researchers and clinicians to make informed decisions about which model to use in their own studies.
- A thorough analysis of hyperparameter tuning methods used in machine learning, providing insights into which methods are the most effective for improving model performance.
- An emphasis on the importance of preprocessing the dataset to ensure optimal model performance, which has been overlooked in previous studies.

Overall, the contributions of this project have the potential to significantly advance the field of disease prognosis and prediction, leading to improved patient outcomes and personalized treatment plans.

1.4 Background knowledge

1.4.1 Machine Learning

Machine learning (ML) is a field of artificial intelligence (AI) that involves teaching computers to learn from data and make predictions or decisions without being explicitly programmed. In other words, it is the process of training a computer system to recognize patterns in data and make intelligent decisions based on those patterns. Machine learning algorithms use statistical techniques to find patterns in large amounts of data and use those patterns to make predictions or decisions about new data. The goal of machine learning is to create algorithms that can learn and improve over time, without human intervention, by continuously analyzing new data and updating their predictions or decisions accordingly. ML has applications in various fields such as image recognition, natural language processing, fraud detection, recommendation systems, and many others.

DECISION TREE

Decision Tree (DT) is a popular machine learning algorithm that can be used for classification tasks such as the Breast Cancer Wisconsin dataset. Here are some reasons why you might choose to use Decision Trees for this dataset:

Interpretable: Decision Trees are easy to understand and interpret. The tree structure allows you to see which features are most important in making the classification decision, making it easier to explain to others and gain insights into the data.

Non-parametric: Decision Trees do not assume any specific distribution or functional form of the data, making them more flexible than parametric models such as logistic regression.

Handles linear or non-linear relationships: Decision Trees can model non-linear relationships between the input variables and the target variable, which may be present in the Breast Cancer Wisconsin dataset.

Scalability: Decision Trees can handle large datasets with many features and can be easily parallelized to speed up training.

High accuracy: Decision Trees can achieve high accuracy on classification tasks, particularly when combined with ensemble methods such as Random Forests.

MATHEMATICAL FORMULATION FOR DECISION TREE

Let X be the feature matrix of shape (n_samples, n_features) where each row corresponds to a sample and each column corresponds to a feature. Let y be the target vector of shape (n_samples,) where each element corresponds to the class label of the corresponding sample (0 for malignant, 1 for benign).
The decision tree algorithm partitions the feature space into regions by recursively splitting the data based on the feature that maximizes the information gain. At each node of the tree, the algorithm selects the feature k and threshold t that maximizes the information gain, defined as:

$$IG(X,\ y,\ k,\ t) = \ H(y)\ -\ H(y \mid X_k\ <\ t)\ -\ H(y \mid X_k\ >=\ t)$$

where $H(y)$ is the entropy of the target variable, and $H(y \mid X_k\ <\ t)$ and $H(y \mid X_k\ >=\ t)$ are the conditional entropies of the target variable given the feature values less than and greater than the threshold t for feature k, respectively.
The splitting process continues until a stopping criterion is met, such as reaching a maximum depth, having a minimum number of samples per leaf, or achieving a minimum improvement in information gain.
Once the decision tree is trained, it can be used to make predictions on new samples by traversing the tree from the root node to a leaf node, where the majority class label of the samples in that leaf node is assigned to the new sample.

SUPPORT VECTOR MACHINE (SVM) LINEAR AND POLY

SVM is a powerful classification algorithm that can be used for both linear and non-linear classification tasks. Here are some reasons why you might choose to use SVM for the Pima Diabetes dataset and cervical cancer dataset:

Margin maximization: SVM tries to find a hyperplane that maximizes the margin between the classes, which helps to reduce the risk of overfitting and improve generalization performance.

Robust to outliers: SVM is less affected by outliers than some other algorithms, such as decision trees.

Flexible: SVM can use different kernel functions to model non-linear relationships between the input variables and the target variable.

MATHEMATICAL FORMULATION FOR SVM LINEAR

Given a training set (X, y) of n samples, where X is the feature matrix of shape (n_samples, n_features), and y is the target vector of shape (n_samples,), SVM linear tries to find a hyperplane that separates the samples into two classes with the largest possible margin. The hyperplane can be represented as $w.x + b = 0$, where w is the weight vector of shape (n_features,), b is the bias scalar, and x is the input feature vector. SVM linear solves the following optimization problem:

minimize $(\frac{1}{2}) \times \sqrt{||w||}$ subject to $y_i (w.x_i + b) \geq 1$ for $i = 1, 2, ..., n$

where $||w||$ is the L2 norm of the weight vector, and y_i is the class label of the i-th sample. Once the hyperplane is found, SVM linear can make predictions on new samples by computing the sign of $w.x + b$.

MATHEMATICAL FORMULATION FOR SVM POLY

SVM poly extends SVM linear to handle non-linear classification tasks by using a polynomial kernel function. The kernel function computes the inner product of the feature vectors in a higher-dimensional space, allowing SVM to find a non-linear decision boundary. SVM poly solves the following optimization problem: minimize $(\frac{1}{2}) \times \sqrt{||w||}$ subject to $y_i(w.phi(x_i) + b) \geq 1$ for i=1,2,...,n where $phi(x_i)$ is the feature vector in the higher-dimensional space, and y_i is the class label of the i-th sample. Once the hyperplane is found, SVMpoly can make predictions on new samples by computing the sign of $w.phi(x_i) + b$.

RANDOM FOREST

Random Forest (RF) is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. Here are some reasons why you might choose to use Random Forests for the Pima Indian Diabetes dataset:

Improved accuracy: Random Forests can achieve higher accuracy than a single decision tree by averaging the predictions of many trees, which helps to reduce overfitting and improve generalization to new samples.

Handles non-linear relationships: Random Forests can model non-linear relationships between the input variables and the target variable, which may be present in the Pima Indian Diabetes dataset.

Variable importance: Random Forests can provide a measure of feature importance, which can help to identify the most relevant features for making the classification decision.

Handles missing values: Random Forests can handle missing values in the dataset without imputing them, which can save time and reduce bias in the analysis.

MATHEMATICAL FORMULATION FOR RANDOM FOREST

Random Forest combines multiple decision trees trained on different subsets of the data and/or features. Each decision tree is trained on a random subset of the data (with replacement) and a random subset of the features. The randomness in the training process helps to reduce overfitting and increase the diversity of the trees. The predictions of the individual trees are combined by majority voting for classification tasks or by averaging for regression tasks. The final prediction is the class label or the mean of the predictions of the individual trees.

LOGISTIC REGRESSION (LR)

Logistic Regression is a popular machine learning algorithm that can be used for classification tasks such as the Pima Indian Diabetes dataset. Here are some reasons why you might choose to use Logistic Regression for this dataset:

Interpretable: Logistic Regression is easy to interpret, allowing you to understand which features are most important in making the classification decision.

Handles linear relationships: Logistic Regression assumes a linear relationship between the input variables and the target variable, which may be sufficient for the Pima Indian Diabetes dataset.

Handles binary classification: Logistic Regression is well-suited for binary classification tasks such as this dataset, where the target variable takes on only two values (0 for no diabetes, 1 for diabetes).

Scalability: Logistic Regression can handle large datasets with many features and can be easily parallelized to speed up training.

Regularization: Logistic Regression can be regularized to prevent overfitting and improve generalization performance.

MATHEMATICAL FORMULATION FOR LOGISTIC REGRESSION

Let X be the feature matrix of shape (n_samples, n_features) where each row corresponds to a sample and each column corresponds to a feature. Let y be the target vector of shape (n_samples,) where each element corresponds to the class label of the corresponding sample (0 for no diabetes, 1 for diabetes).

The logistic regression algorithm models the probability of belonging to the positive class (diabetes) as a function of the input features. Specifically, it models the log odds ratio, or logit, as a linear function of the input features:

$$logit(p) = log(\frac{p}{(1-p)}) = w_0 + w_1 \times X_1 + \ldots + w_p \times X_p$$

where p is the probability of belonging to the positive class, w_0 is the intercept term, and w_1, ..., w_p are the coefficients for the input features X_1, ..., X_p.

The logistic function, or sigmoid, is used to convert the logit to a probability:

$$p = \frac{1}{(1 + e^{(-logit(p))})}$$

The logistic regression algorithm learns the optimal values of the coefficients w_1, ..., w_p by minimizing the binary cross-entropy loss between the predicted probabilities and the true class labels:

$$L(w) = -(\frac{1}{n}) \times \sum_{i=0}^{p} (y_i \times log(p_i) + (1 - y_i) \times log(1 - p_i))$$

where n is the number of samples, y_i is the true class label of sample i, and p_i is the predicted probability of sample i belonging to the positive class.

The optimization problem can be solved using gradient descent or other optimization algorithms.

In summary, Logistic Regression is a simple yet effective algorithm for binary classification tasks such as the Pima Indian Diabetes dataset. Its interpretable nature and ability to handle linear relationships make it a popular choice in many applications.

1.4.2 Description of Evaluation Parameter

1. Precision or Positive Predictive Value (PPV): It measures the proportion of predicted positive instances that are actually positive. It is useful when the cost of false positives is high, and we want to minimize them.

2. Recall or Sensitivity or True Positive Rate (TPR) or Hit Rate: It measures the proportion of actual positive instances that are correctly identified by the model. It is useful when the cost of false negatives is high, and we want to minimize them.

3. F1 Score or Harmonic Mean: It is the weighted average of precision and recall, and it balances both the measures. It is useful when we want to find a balance between precision and recall.

4. F1 Measure: It is the arithmetic mean of precision and recall, and it is similar to the F1 score. It is useful when we want a simpler measure than the F1 score.

5. Specificity or True Negative Rate (TNR) or Selectivity: It measures the proportion of actual negative instances that are correctly identified by the model. It is useful when the cost of false positives is high, and we want to minimize them.

6. Negative Predictive Value (NPV): It measures the proportion of predicted negative instances that are actually negative. It is useful when the cost of false negatives is high, and we want to minimize them.

7. False Positive Rate (FPR): It measures the proportion of actual negative instances that are incorrectly identified as positive by the model. It is useful when the cost of false positives is high, and we want to minimize them.

8. False Negative Rate (FNR) or Miss Rate: It measures the proportion of actual positive instances that are incorrectly identified as negative by the model. It is useful when the cost of false negatives is high, and we want to minimize them.

9. False Discovery Rate (FDR): It measures the proportion of predicted positive instances that are actually negative. It is useful when the cost of false positives is high, and we want to minimize them.

10. Critical Success Index or Threat Score (CSI): It measures the proportion of correctly identified positive instances out of the total number of positive instances. It is useful when the cost of false negatives is high, and we want to minimize them.

11. Fowlkes-Mallows Index (FM): It is a geometric mean of precision and recall, and it is useful when we want to balance precision and recall.

12. Balanced Accuracy (BA): It is the average of sensitivity and specificity, and it is useful when the dataset is imbalanced.

13. Matthews Correlation Coefficient (MCC): It is a measure of the quality of binary (two-class) classifications, and it is useful when the dataset is imbalanced.

14. Bookmaker Informedness or Informedness (BI): It measures the proportion of correct decisions made by the model, and it is useful when we want to minimize the number of incorrect decisions.

15. Markedness or delta (MK): It measures the proportion of correct predictions made by the model, and it is useful when we want to minimize the number of incorrect predictions.

16. False Omission Rate (FOR): It measures the proportion of actual negative instances that are incorrectly identified as negative by the model. It is useful when the cost of false negatives is high, and we want to minimize them.

17. Positive Likelihood Ratio (PLR): It measures the likelihood of a positive prediction being correct, and it is useful when we want to minimize the number of incorrect positive predictions.

18. Negative Likelihood Ratio (NLR): It measures the likelihood of a negative prediction being correct, and it is useful when we want to minimize the number of incorrect negative predictions.

19. Prevalence Threshold (PT) is a metric that helps to determine the minimum probability threshold required for a positive prediction to be considered reliable. It takes into account the balance between the false positive rate (FPR) and the true positive rate (Recall) and is particularly useful in situations where the prevalence of the positive class is low. A higher PT value indicates that the positive class should be predicted with a higher probability threshold, and a lower PT value indicates a lower threshold.

20. Diagnostic Odds Ratio (DOR) is a metric that measures the effectiveness of a diagnostic test. It is the ratio of the odds of a positive test result in people with the condition to the odds of a positive test result in people without the condition. DOR ranges from 0 to infinity, with

higher values indicating better discriminatory power of the test. DOR can be used to compare the diagnostic accuracy of different tests and to determine the optimal cutoff value for a test.

21. Accuracy: The accuracy is the ratio of the number of correct predictions to the total number of predictions made by the model. It provides an overall measure of the model's performance and is useful when the classes are balanced. However, accuracy can be misleading when the classes are imbalanced, and the model may perform well on the majority class but poorly on the minority class.

22. Cohen's Kappa Score: Cohen's Kappa is a statistic that measures the agreement between the predicted and actual class labels, taking into account the possibility of chance agreement. It is often used when the classes are imbalanced or when the distribution of the predicted and actual class labels is different. A high Kappa score indicates good agreement between the predicted and actual class labels, while a low Kappa score indicates poor agreement. It is generally interpreted as follows: less than 0 is indicating no agreement, 0.01 to 0.20 as slight agreement, 0.21 to 0.40 as fair agreement, 0.41 to 0.60 as moderate agreement, 0.61 to 0.80 as substantial agreement, and 0.81 to 1 as almost perfect agreement.

Overall, these metrics provide valuable information on the performance of a classification model and can help in selecting the best model for a given problem. However, it is important to consider the nature of the problem, the characteristics of the dataset, and the class distribution when interpreting the results of these metrics.

1.5 Review of Literatures

Table 1 :-> Literature survey

Reference	Objective	Contribution	Limitations
Youcef Djenouri, et.al [1]	With the growth of smart medical devices and applications in the smart hospitals, home care facilities, nursing, and the Internet of Medical Things are becoming more ubiquitous and to detect key body indicators, monitor health situations, and generate multivariate data to provide just-in-time healthcare services.	In this article they present a novel collaborative disease detection system based on IoMT amalgamated with captured image data.	The integration of smart medical devices and applications with captured image data has led to the development of a collaborative disease detection system based on the Internet of Medical Things. However, the study has limitations such as the use of a single data source and ethical considerations, and the technical implementation can be challenging. Furthermore, the limited evaluation of the system on external datasets and different settings may limit its generalizability, and the clinical relevance of the system requires further investigation.

Tejas N. Joshi, Prof. Pramila M. Chawan [3]	the paper proposes a medical Decision Support System (DSS) for diabetes prediction based on Machine Learning (ML) techniques, and compares the performance of conventional machine learning with deep learning approaches. The study finds that the Random Forest classifier is more effective for diabetes prediction compared to the deep learning and SVM methods.	the project aims to propose an effective method for earlier detection of diabetes.	While the proposed Medical Decision Support System (DSS) for diabetes prediction based on Machine Learning (ML) techniques shows promise, there are limitations to the study. One limitation is the reliance on a single dataset, which may not be representative of the broader population. Furthermore, the study does not address the cost-effectiveness of the proposed system, which may limit its adoption in healthcare settings. Additionally, the study only focuses on comparing conventional machine learning with deep learning approaches and does not consider other emerging techniques, which may yield better results. Moreover, the study does not address potential ethical considerations such as patient privacy and informed consent. Finally, the study does not evaluate the potential impact of the proposed system on patient outcomes, which is crucial for determining the system's clinical relevance.

Aishwarya Mujumdar , V Vaidehi Dr. [2]	The main goal of this research paper is to introduce a diabetes prediction model that takes into account not only regular factors but also external factors to enhance the accuracy of diabetes classification. The paper endeavors to demonstrate that the inclusion of external factors, such as occupation, income, and family history, can considerably improve the accuracy of diabetes prediction models.	In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with the new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.	The research paper proposes a diabetes prediction model that incorporates external factors, such as occupation and income, to improve classification accuracy. However, the study only considers a limited number of external factors and does not address potential ethical considerations such as patient privacy and informed consent. The study also does not evaluate the potential impact of the proposed model on patient outcomes or address the cost-effectiveness of the model.
Amani Yahyaoui; Akhtar Jamil; Jawad Rasheed; Mirsat Yesiltepe [6]	The objective of the research paper is to propose and evaluate the effectiveness of a medical Decision Support System (DSS) for diabetes prediction using Machine Learning (ML) techniques. The paper aims to compare the performance of conventional machine learning (using Support Vector Machine and Random Forest classifiers) with deep learning (using a fully Convolutional Neural Network) approaches for predicting and detecting diabetes patients	The paper proposes a medical Decision Support System (DSS) for diabetes prediction based on Machine Learning (ML) techniques, and compares the performance of conventional machine learning with deep learning approaches. The study finds that the Random Forest classifier is more effective for diabetes prediction compared to the deep learning and SVM methods.	The study's limitations include the lack of external validation on different datasets, which may limit the generalizability of the results. The study also does not address ethical considerations such as patient privacy and informed consent. Additionally, the cost-effectiveness of the proposed system is not evaluated, limiting its potential adoption in healthcare settings.

Debadri Dutta; Debpriyo Paul; Parthajeet Ghosh [5]	The objective of this paper is to identify the key factors that contribute to the development of diabetes, specifically those related to obesity and high blood sugar levels. The paper aims to focus on variable and feature selection in order to determine the most important features for predicting the likelihood of developing diabetes in the future. Overall, the research aims to provide insights into the factors that are most significant in the development of diabetes, and to inform future research on the topic of diabetes prediction and prevention.	the paper investigates the key factors that contribute to the development of diabetes, particularly those related to obesity and high blood sugar levels. The paper emphasizes the importance of variable and feature selection in identifying the most significant features for predicting the likelihood of developing diabetes in the future. The study suggests that obesity and high blood sugar levels are major risk factors for diabetes, and that these factors should be considered in future research on diabetes prediction and prevention.	One limitation of this paper is that it focuses solely on the relationship between obesity, high blood sugar levels, and the development of diabetes, and does not consider other potential risk factors. Additionally, the study does not address potential ethical considerations related to the use of personal health data for predictive purposes.

Muhammad Azeem Sarwar; Nasir Kamal; Wajeeha Hamid; Munam Ali Shah [4]	The objective of this paper is to explore the use of predictive analytics in healthcare, specifically in the context of diabetes prediction. The paper discusses the application of six different machine learning algorithms on a dataset of patient medical records, with the goal of identifying the algorithm that is best suited for predicting diabetes. The research aims to compare the performance and accuracy of these different machine learning techniques, and to provide insights into the potential of predictive analytics to help doctors and practitioners make timely and informed decisions about patient health and treatment.	The research demonstrates the potential of machine learning techniques to improve patient outcomes by enabling timely and informed decisions about health and treatment. The study applies six different machine learning algorithms to a dataset of patient medical records, and compares their performance and accuracy for diabetes prediction. The results of the study provide insights into which algorithm is best suited for this task, and can help inform the development of predictive analytics tools for healthcare practitioners.	One limitation of this paper is the use of a single dataset for analysis, which may not be representative of the broader population. Additionally, the study does not consider potential ethical concerns, such as patient privacy and informed consent, which are important considerations in the use of predictive analytics in healthcare.

S. Saru, S. Subhasree [7]	The objective of this research paper is to investigate the application of data mining techniques to predict and diagnose diabetes using the Pima Indian diabetes database. The study aims to evaluate the performance of different classification algorithms, including Naïve Bayes, Decision Trees, and KNN, in diagnosing diabetes. Additionally, the research aims to apply the bootstrapping resampling technique to enhance the accuracy of the classification models. The goal is to identify valuable patterns in the dataset that can provide significant information to medical professionals in diagnosing and managing diabetes, a disease that can lead to serious health complications. Ultimately, the paper aims to contribute to the development of an effective prediction system for diabetes that can improve healthcare outcomes and reduce the burden of this disease on patients and healthcare systems.	The research demonstrates that data mining techniques can be effective in predicting and diagnosing diabetes using the Pima Indian diabetes database. The study shows that classification algorithms such as Naïve Bayes, Decision Trees, and KNN can be applied to analyze the dataset and identify valuable patterns for diagnosing diabetes. Additionally, the bootstrapping resampling technique can be used to enhance the accuracy of the classification models. Overall, the research paper demonstrates the potential of data mining techniques in healthcare, specifically in improving the management and diagnosis of diabetes, a disease that can lead to serious health complications.	The paper identifies several limitations in its application of data mining techniques to predict and diagnose diabetes using the Pima Indian diabetes database. Firstly, the study focuses only on the Pima Indian population, which may not represent the broader population. Thus, the findings may not be generalizable to other populations. Secondly, the accuracy and reliability of the data used in the study are crucial factors that may affect the performance of data mining techniques. Thirdly, the study only evaluates the performance of three classification algorithms, which may not be enough to provide a comprehensive comparison of different data mining techniques. Fourthly, the lack of external validation of the classification models may affect the generalization of the results. Additionally, ethical considerations such as patient privacy and informed consent are not addressed in the study, which may raise concerns for medical research. Lastly, although the study shows the potential of data mining techniques in predicting and diagnosing diabetes, it does not provide clear clinical implications for healthcare providers or patients.

Deepti Sisodia , Dilip Singh Sisodia [8]	The objective of this research paper is to design a model for predicting the likelihood of diabetes in patients using machine learning classification algorithms. The study aims to evaluate the performance of three algorithms, namely Decision Tree, SVM, and Naive Bayes, in detecting diabetes at an early stage.	This article demonstrates that machine learning classification algorithms, such as Decision Tree, SVM, and Naive Bayes, can be effective in predicting the likelihood of diabetes in patients. The study shows that Naive Bayes outperforms other algorithms with the highest accuracy of 76.30%. The experiments were conducted on the Pima Indians Diabetes Database (PIDD) sourced from the UCI machine learning repository, and the performance of the algorithms was evaluated based on measures like Precision, Accuracy, F-Measure, and Recall. The results of the study demonstrate the potential of machine learning approaches in healthcare, specifically in improving the detection and diagnosis of diabetes at an early stage, which can lead to better patient outcomes and reduced healthcare costs.	Insufficient Feature Selection: The study did not include a detailed analysis of feature selection, which is crucial in the performance of machine learning algorithms. The performance of the algorithms may be improved with better feature selection. Lack of External Validation: The study did not validate the performance of the algorithms on external datasets, which may affect the generalization of the results. Incomplete Evaluation: The study only evaluated the performance of three algorithms, which may not be sufficient to provide a comprehensive comparison of different machine learning techniques. Ethical Considerations: The study did not address ethical considerations, such as patient privacy and informed consent, which are critical issues in medical research.

Md Abu Rumman Refat; Md. Al Amin; Chetna Kaushal; Mst Nilufa Yeasmin; Md Khairul Islam [9]	The objective of this research paper is to conduct a comparative analysis of several machine learning and deep learning techniques for early diabetes disease prediction. The study aims to evaluate the performance of these techniques using a diabetes dataset from the UCI repository, which includes 17 attributes, including class. Ultimately, the paper aims to contribute to the development of an accurate and efficient model for diagnosing diabetes that can help improve patient outcomes and reduce the burden of the disease on healthcare systems.	This research illustrates that machine learning and deep learning techniques can effectively predict the onset of diabetes at an early stage, which is crucial for the timely treatment of the disease. The study shows that several machine learning and deep learning algorithms, including XGBoost, can achieve high levels of accuracy in predicting diabetes.	One limitation of this study is that it relies on a single dataset, which may not be representative of the broader population. The study also does not address potential ethical considerations, such as patient privacy and informed consent. Additionally, the paper only focuses on evaluating the performance of several machine learning and deep learning techniques, without considering other emerging techniques that may yield better results. Finally, the study does not evaluate the potential impact of the proposed model on patient outcomes, which is crucial for determining the clinical relevance of the model in real-world settings.

N. Yuvaraj & K. R. SriPreethaa [10]	The objective of this research paper is to propose a novel implementation of machine learning algorithms in a Hadoop-based cluster for diabetes prediction. The paper aims to demonstrate how the use of machine learning algorithms in a distributed computing environment can improve the accuracy of disease prediction in healthcare systems. The study uses the Pima Indians Diabetes Database to evaluate the effectiveness of the proposed approach. The ultimate goal of this research is to contribute to the improvement of healthcare systems and to enhance disease prediction accuracy using advanced computing techniques.	In summary, the paper suggests that the use of machine learning algorithms in Hadoop-based clusters can lead to the development of highly accurate healthcare systems for predicting diabetes. Nonetheless, further investigations are required to assess the efficacy of this methodology in diverse datasets and practical settings.	While the proposed implementation of machine learning algorithms in a Hadoop-based cluster for diabetes prediction is promising, there are some limitations to the study. The study only focuses on the Pima Indians Diabetes Database and does not consider other datasets, limiting the generalizability of the results. Additionally, the study does not address the scalability of the proposed system, which may be challenging to implement in larger healthcare systems. Furthermore, the technical complexity of implementing a Hadoop-based cluster may limit the adoption of this approach in some healthcare settings. Finally, the study does not evaluate the potential impact of the proposed system on patient outcomes, which is crucial for determining the clinical relevance of the system.

Jobeda Jamal Khanam, Simon Y. Foo [11]	The objective of this research paper is to evaluate the performance of different data mining, machine learning, and neural network methods for diabetes prediction using the Pima Indian Diabetes dataset. The study aims to determine which algorithm can accurately predict diabetes in patients at an early stage, with a focus on Logistic Regression, Support Vector Machine, and Neural Network models. The research also aims to compare the performance of these algorithms and determine the optimal number of hidden layers and epochs for the Neural Network model.	This research paper demonstrates that machine learning algorithms, including Logistic Regression and Support Vector Machine, can be effective in predicting diabetes using the Pima Indian Diabetes dataset. Additionally, the research found that a Neural Network model with two hidden layers can achieve 88.6% accuracy. However, further research may be necessary to evaluate the effectiveness of these approaches in other datasets and real-world scenarios.	One limitation of this research paper is the reliance on a single dataset, which may not be representative of other populations or healthcare settings. Additionally, the study does not address potential ethical considerations, such as patient privacy and informed consent, which could impact the adoption and implementation of these methods in clinical practice. Another limitation is that the research only considers a limited number of machine learning algorithms and neural network models, which may not capture the full range of potential approaches. Finally, the study does not evaluate the clinical relevance of these prediction models, such as their impact on patient outcomes or healthcare costs, which is crucial for determining their practical utility.

Safial Islam Ayon, Md. Milon Islam [12]	The objective of this research paper is to propose a strategy for the diagnosis of diabetes using deep neural network by training its attributes in five-fold and ten-fold cross-validation fashion. The study aims to evaluate the effectiveness of the proposed system on the Pima Indian Diabetes dataset and to demonstrate that deep learning approach can be an auspicious system for the prediction of diabetes with high accuracy. The research also aims to investigate the performance of the proposed system in terms of prediction accuracy, F1 score, and MCC for five-fold cross-validation, and to evaluate the accuracy, sensitivity, and specificity for ten-fold cross-validation. The experimental results are expected to demonstrate the promising performance of the proposed system in predicting diabetes, especially for five-fold cross-validation.	This paper demonstrates that using a deep neural network approach for the diagnosis of diabetes can provide a promising and effective system for predicting diabetes with high accuracy. The research used the Pima Indian Diabetes dataset and achieved a prediction accuracy of 98.35% with five-fold cross-validation and 97.11% with ten-fold cross-validation. The results suggest that the proposed system can be a useful tool for diagnosing diabetes and could potentially be applied in real-world scenarios. However, further research is needed to evaluate the effectiveness of this approach with other datasets and in different settings.	One limitation of this research paper is the use of only one dataset, the Pima Indian Diabetes dataset. Although this dataset is widely used in diabetes prediction studies, it may not be representative of other populations or ethnic groups. Additionally, the study does not investigate the interpretability of the deep neural network model or provide insights into which attributes or features are most important for predicting diabetes. Lastly, the study does not address the feasibility or scalability of deploying the proposed system in real-world clinical settings, which may have practical limitations such as data privacy concerns and computational resources.

Jayroop Ramesh, Raafat Aburukba, Assim Sagahyroon [13]	The objective of this research paper is to propose an end-to-end remote monitoring framework for automated diabetes risk prediction and management using personal health devices, smart wearables and smartphones. The study aims to develop a support vector machine for diabetes risk prediction using the Pima Indian Diabetes Database and achieve competitive performance metrics. The framework also aims to enable medical professionals to make informed decisions based on the latest diabetes risk predictions and lifestyle insights while attaining unobtrusiveness, reduced cost, and vendor interoperability.	The research paper illustrates that an end-to-end remote monitoring framework can be developed for automated diabetes risk prediction and management using personal health devices, smart wearables, and smartphones. The proposed support vector machine for diabetes risk prediction achieved competitive performance metrics, including accuracy, sensitivity, and specificity scores. The framework enables medical professionals to make informed decisions based on the latest diabetes risk predictions and lifestyle insights while maintaining unobtrusiveness, reduced cost, and vendor interoperability.	One limitation of this research paper is the potential bias and generalizability issues associated with using the Pima Indian Diabetes Database, which may not be representative of other populations or ethnic groups. Additionally, the study does not investigate the interpretability of the support vector machine model or provide insights into which features or variables are most important for diabetes risk prediction. Another limitation is the lack of consideration for potential ethical and privacy concerns associated with remote monitoring of personal health data. Lastly, the study does not address the feasibility or scalability of implementing the proposed framework in real-world clinical settings, which may have practical limitations such as data privacy concerns and technical infrastructure requirements.

Ioannis Kavakiotis, Olga Tsave , Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvard [14]	The objective of this research paper is to conduct a systematic review of the applications of machine learning and data mining techniques in diabetes research. The paper aims to categorize the applications into four main areas: prediction and diagnosis, diabetic complications, genetic background and environment, and health care and management. The paper also aims to analyze the machine learning algorithms used and the type of data employed in the selected articles. The overall goal of the paper is to highlight the usefulness of these techniques in extracting valuable knowledge and generating new hypotheses for further investigation in the field of diabetes.	The paper highlights the importance of utilizing advanced computational techniques in the field of diabetes research to make sense of the vast amount of available data and improve patient outcomes.	One limitation of this research paper is the potential for publication bias, as the selected articles may not represent the full range of studies on the applications of machine learning and data mining techniques in diabetes research. Additionally, the paper may not capture the latest developments in the field, as the review is limited to studies published up to a certain date. Another limitation is the possibility of heterogeneity in the studies included, such as differences in the populations studied, outcome measures, and machine learning algorithms used, which may affect the generalizability of the findings. Lastly, the paper does not provide insights into the interpretability or explainability of the machine learning models used, which may limit their clinical utility.

Huma Naz & Sachin Ahuja [15]	The International Diabetes Federation (IDF) stated that 382 million people are living with diabetes worldwide. Over the last few years, the impact of diabetes has increased drastically, which makes it a global threat. At present, Diabetes has steadily been listed in the top position as a major cause of death. The number of affected people will reach up to 629 million i.e. 48% increase by 2045. However, diabetes is largely preventable and can be avoided by making lifestyle changes. These changes can also lower the chances of developing heart disease and cancer. So, there is a dire need for a prognosis tool that can help the doctors with early detection of the disease and hence can recommend the lifestyle changes required to stop the progression of the deadly disease.	The outcome of the study confirms that DL provides the best results with the most promising extracted features. DL achieves the accuracy of 98.07% which can be used for further development of the automatic prognosis tool. The accuracy of the DL approach can further be enhanced by including the omics data for prediction of the onset of the disease.	One limitation of this statement is the lack of specificity regarding the types of diabetes being referred to, as there are several subtypes with varying causes, risk factors, and treatment approaches. Additionally, the statement does not provide insights into the regional or demographic disparities in diabetes prevalence and burden, which may have important implications for public health policies and interventions. Another limitation is the potential oversimplification of the notion that lifestyle changes alone can prevent or manage diabetes, as genetic, environmental, and other factors may also play a significant role. Lastly, the statement does not address the potential barriers to accessing diagnosis and treatment for diabetes, particularly in low-resource settings or disadvantaged communities.

Souad Larabi-Marie-Saint e, ORCID,Linah Aburahmah, Rana Almohaini and Tanzila Saba [17]	The objective of this research paper is to survey the Machine Learning (ML) and Deep Learning (DL) techniques used for predicting diabetes in the last six years and compare their performance. Additionally, the study aims to analyze the performance of rarely or not used ML classifiers in diabetes prediction and recommend their use in combination with other techniques for enhancing prediction accuracy. The research seeks to contribute to the development of more robust and accurate models for diabetes prediction using a wider range of classifiers.	This paper provides a comprehensive survey of ML and DL techniques used in diabetes prediction and examines the performance of rarely used classifiers on the Pima Indian dataset. The results indicate that these classifiers can achieve an accuracy of 68%–74%, suggesting that they have potential for further development and can be used in combination with other techniques to improve diabetes prediction. The study underscores the importance of continuing to explore and develop different ML and DL techniques to enhance diabetes prediction.	One limitation of this research paper is the potential for publication bias, as the study may not include all relevant studies on the topic, particularly those published outside the selected time frame or in non-indexed or non-English language journals. Additionally, the study may not account for variations in the datasets used in the selected studies, which may affect the generalizability of the findings. Another limitation is the lack of consideration for potential ethical and privacy concerns associated with the use of personal health data in ML and DL models for diabetes prediction. Lastly, the study does not address the feasibility or scalability of implementing the recommended classifiers in real-world clinical settings, which may have practical limitations such as data privacy concerns and technical infrastructure requirements.

Md. Maniruzzaman, Md. Jahanur Rahman, Benojir Ahammed & Md. Menhazul Abedin [16]	Diabetes is a chronic disease characterized by high blood sugar. It may cause many complicated disease like stroke, kidney failure, heart attack, etc. About 422 million people were affected by diabetes disease in worldwide in 2014. The figure will be reached 642 million in 2040. The main objective of this study is to develop a machine learning (ML)-based system for predicting diabetic patients.	The combination of LR and RF-based classifier performs better. This combination will be very helpful for predicting diabetic patients.	One limitation of this study is the potential bias in the dataset used for developing the ML-based system, which may not be representative of the broader population of diabetic patients. Additionally, the study may not account for the dynamic and complex nature of diabetes progression and management, as the ML-based system may not adapt well to changes in patient conditions or treatment responses over time. Another limitation is the lack of consideration for the interpretability and explainability of the ML-based system, which may limit its clinical utility and ethical implications. Lastly, the study does not address the potential challenges and limitations associated with implementing the ML-based system in real-world clinical settings, such as technical infrastructure requirements and regulatory compliance.

Akm Ashiquzzaman, Abdul Kawsar Tushar, Md. Rashedul Islam, Dongkoo Shon, Kichang Im, Jeong-Ho Park, Dong-Sun Lim & Jongmyon Kim [18]	The objective of this research paper is to propose a reliable prediction system for the disease of diabetes using a dropout method to address the issue of overfitting. The paper presents the use of deep learning neural network, where fully connected layers are followed by dropout layers, to improve the prediction accuracy of the Pima Indians Diabetes Data Set. The main focus of the paper is to highlight the effectiveness of the proposed method in outperforming other state-of-the-art techniques in predicting diabetes accurately.	deep learning neural networks with dropout layers have been proposed as a reliable method for predicting diabetes, particularly in addressing the issue of overfitting. This paper provides evidence that the proposed method outperforms other state-of-the-art methods in terms of prediction accuracy for the Pima Indians Diabetes Data Set. The findings suggest that this method may be a valuable tool for improving diabetes prognosis.	One limitation of this research paper is that the proposed method may not be easily generalizable to other datasets or patient populations, as the study focuses specifically on the Pima Indians Diabetes Data Set. Additionally, the study may not account for potential biases or confounding factors in the dataset, which may affect the generalizability of the results. Another limitation is the lack of consideration for the interpretability and explainability of the deep learning neural network, which may limit its clinical utility and ethical implications. Lastly, the study does not address the potential challenges and limitations associated with implementing the proposed method in real-world clinical settings, such as technical infrastructure requirements and regulatory compliance.

Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh & Gregor Stiglic [19]	The objective of this research paper is to compare machine learning-based prediction models to commonly used regression models for the prediction of undiagnosed T2DM. The study aims to measure the performance of different models in predicting fasting plasma glucose levels using 100 bootstrap iterations in different subsets of data simulating new incoming data in 6-month batches. The research also evaluates the stability of variable selection over time for each model and considers the interpretability and model calibration in the development of clinical prediction models. Overall, the study aims to provide insights into the effectiveness of machine learning approaches in developing accurate and stable prediction models for T2DM.	The research paper reveals that the use of simpler regression models can be as effective as more complex machine learning-based models in predicting fasting plasma glucose level for undiagnosed T2DM. The study emphasizes the importance of interpretability and model calibration in the development of clinical prediction models, and notes that greater stability in the selection of variables over time can facilitate simpler interpretation of the models. The paper ultimately suggests that sophisticated prediction models may not necessarily yield significant improvements in the prediction of T2DM.	One limitation of this research paper is that the study focuses only on the prediction of fasting plasma glucose levels, which is only one aspect of T2DM diagnosis. Other important factors, such as HbA1c levels and symptoms, are not considered. Additionally, the study uses simulated incoming data in 6-month batches, which may not accurately represent real-world clinical settings. The study also does not account for potential biases or confounding factors in the dataset, which may affect the generalizability of the results. Lastly, the study does not address the potential challenges and limitations associated with implementing the developed prediction models in real-world clinical settings, such as technical infrastructure requirements and regulatory compliance.

Minyechil Alehegn, Rahul Joshi& Dr. Preeti Mulay [20]	The objective of this research paper is to investigate the effectiveness of various data mining techniques, including SVM, Naïve Net, DecisionStump, and a proposed ensemble method (PEM), in clustering and predicting symptoms of diabetes using the Pima Indian Diabetes Data Set. The proposed PEM is a hybrid model that combines multiple techniques into one for improved accuracy. The goal of this study is to determine the most effective approach for predicting diabetes at an early stage and potentially saving lives.	This paper highlights the potential of ensemble methods in combining individual techniques to improve prediction accuracy in medical datasets. Overall, the paper demonstrates the value of applying data mining techniques in healthcare and emphasizes the need for continued research in this area.	The limitation of this research paper is that it only focuses on the Pima Indian Diabetes Data Set, which may not represent the general population's characteristics. Additionally, the study does not consider external factors such as environmental or lifestyle factors that may affect diabetes prediction accuracy. Finally, the paper does not provide a comprehensive explanation of the proposed ensemble method, making it difficult to understand the method's validity and potential for future use.

1.6 Proposed Solution

The proposed solutions for the problems identified in the project are as follows:

- **Lack of comparison between models:** To address this issue, we have evaluated and compared multiple machine learning models, including decision trees, logistic regression, support vector machines and random forests. By analyzing the performance of these models on the dataset, we can determine the most effective model for disease prognosis and prediction.
- **Lack of comparison and contrast between hyperparameter tuning methods:** To tackle this problem, we have evaluated and compared multiple hyperparameter tuning methods for each of the machine learning models considered. This includes grid search, random search, and Bayesian optimization. By analyzing the performance of these methods on the dataset, we can determine the most effective method for hyperparameter tuning.
- **Minimal preprocessing of the dataset:** To improve model performance, we have conducted extensive preprocessing of the dataset. This includes data cleaning, feature selection, normalization, and balancing the class distribution. By optimizing the dataset for the machine learning models, we can ensure that the models are trained on the most relevant and representative data, leading to more accurate disease prognosis and prediction.

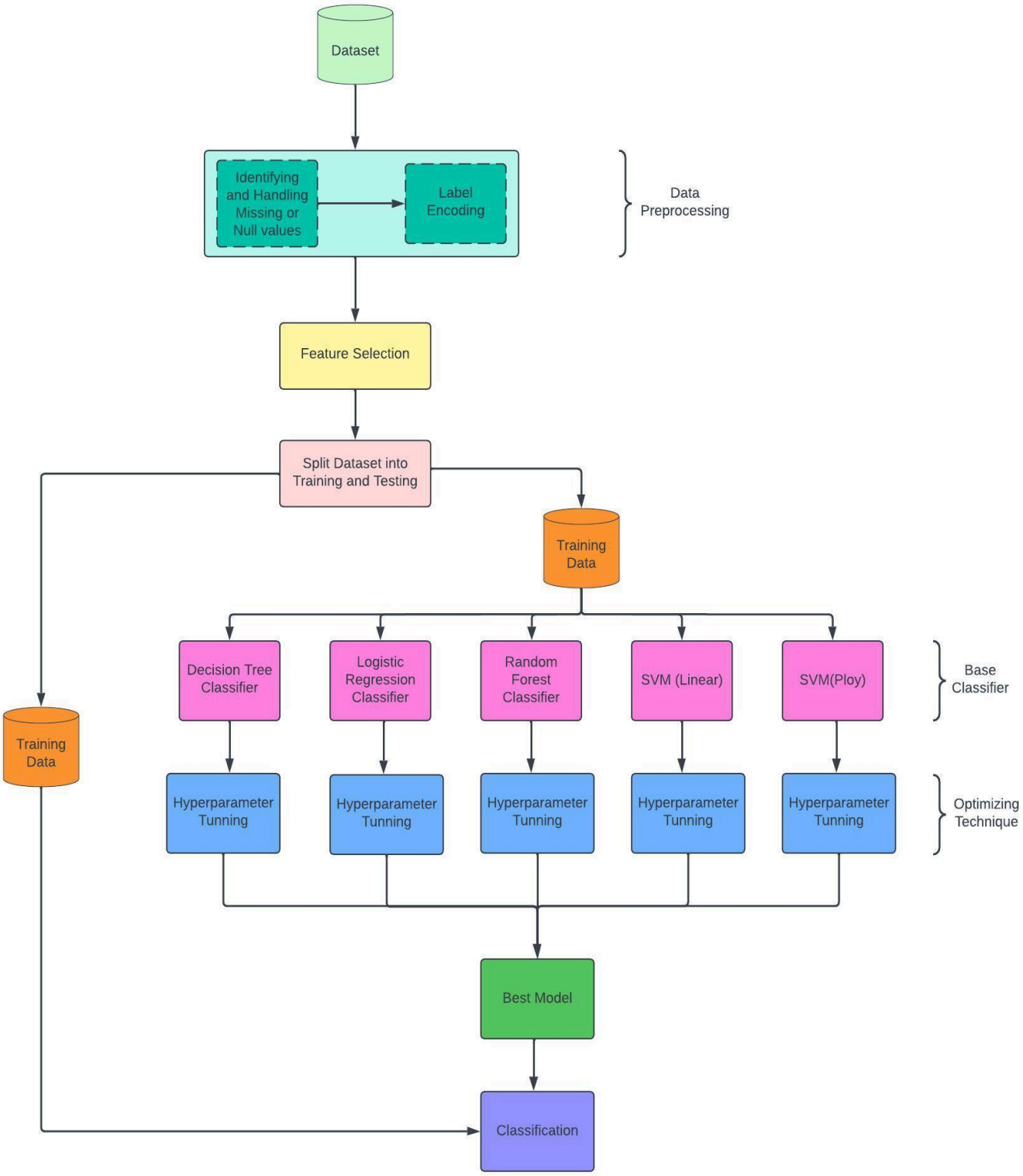
By implementing these proposed solutions, we aim to improve the accuracy and reliability of disease prognosis and prediction models, ultimately leading to better patient outcomes.

Chapter 2 Proposed Research Approach

2.1 Proposed Research Methodology

The proposed algorithm can be applied to develop a predictive model for health conditions using Diabetes [1] and Cervical Cancer datasets [2]. The data is preprocessed by removing missing values, handling outliers, and normalizing or standardizing the features. Feature selection techniques are used to select informative features to reduce model complexity and improve performance. Different machine learning models are trained on the selected features and evaluated using appropriate metrics like accuracy, precision, recall, and F1-score. Hyperparameter tuning techniques are used to find the best hyperparameters for the chosen models. The best model is selected based on its performance and used to make predictions on new data.

2.2 Proposed Block Diagram



2.3 Proposed Algorithm

In this project we approaches this Algorithm

Proposed Algorithm : Diaseases_proganosis_System(D)

Input: Dataset D

Output: Trained best model

1. BEGIN
2. a := Load_Preprocess(D);
3. b := Select_Best_Split(a);
4. c := Best_Feature_Selection(b);
5. d := Select_Best_ModeMachine_Learning_Model(c);
6. e := Find_Best_Hyperparameter(d);
7. f := Choose_Best_Model_Evaluate_Performance(e);
8. return e;
9. **END**

Proposed Algorithm: Load_Preprocess(D)

1. **Input:** Dataset D
2. **Output**: Preprocessed dataset D
3. Preprocessing Method: Removing missing values followed by outlier handling and conversion of categorical features into numerical features using label encoding and subsequent normalization of numerical features using min-max or standard scaler
4. **BEGIN**
5. m := Load(D);
6. n := Preprocess(m);
7. return n;
8. **END**

Proposed Algorithm: Select_Best_Split(a)

1. **Input:** Preprocessed Dataset n
2. **Output:** Best-splitted dataset

3. **BEGIN**
4. Set D1 := Split_Ratio(90:10);
5. Set D2 := Split_Ratio(85:15);
6. Set D3 := Split_Ratio(80:20);
7. Set D4 := Split_Ratio(75:25);
8. Set R1 := Machine_Learning_Models(D1)
9. Set R2 := Machine_Learning_Models(D2)
10. Set R3 := Machine_Learning_Models(D3)
11. Set R4 := Machine_Learning_Models(D4)
12. s := Find_Best_Dataset(R1,R2,R3,R4);
13. return s;
14. **END**

Proposed Algorithm: Select_Best_Feature_Selection(b)

Input: Best-splitted dataset s

Output: Best Feature Selection

1. **BEGIN**
2. Set FS_tech1 := k-best;
3. Set FS_tech2 := mutual info classif;
4. Set FS_tech3 := chi-square;
5. Set FS_tech4 := correlation matrix;
6. t := Choose_Best_Feature_Selection (FS_tech1, FS_tech2, FS_tech3, FS_tech3);
7. return t;
8. **END**

Proposed Algorithm: Select_Best_Machine_Learning_Model(c)

Input : dataset with selected dataset t

Output: best 4 machine learning models

1. **BEGIN**
2. Set ML_Model1 := Decision_Tree(t);
3. Set ML_Model2 := Random_Forest(t);
4. Set ML_Model3 := Logistic_Regression(t);
5. Set ML_Model4 := SVM_linear(t);
6. Set ML_Model5 := SVM_poly(t);
7. Set ML_Model6 := SVM_rbf(t);

8. Set ML_Model7 := SVM_precomputed(t);
9. M:=Choose_Best_4_Machine_Learning(ML_Model1, ML_Model2, ML_Model3, ML_Model4, ML_Model5, ML_Model6, ML_Model7);
10. Return m;
11. **END**

Proposed Algorithm: Find_Best_Hyperparameter(d)

1. **Input:** Best-model t
2. **Output:** Best hyperparameter
3. **BEGIN**
4. Set hyper_tech1 := randomized search(m);
5. Set hyper_tech2 := grid search(m);
6. Set hyper_tech3 := hyperopt(m);
7. Set hyper_tech4 := bayesian optimization(m);
8. Set hyper_tech5 := gradient-based optimization(m);
9. u := Select_Best_Hyperparameters(hyper_tech1, hyper_tech2, hyper_tech3, hyper_tech4, hyper_tech5);
10. return u;
11. **END**

Proposed Algorithm: Choose_Best_Model_Evaluate_Performance(e)

Input: Best-hyperparameters u

Output: Final best model using classical centralized ML algorithm

1. Evaluation Metrices: Accuracy, precision, recall, F1-score.
2. **BEGIN**
3. acc := Find_Best_Accuracy();
4. pre := Find_Best_Precision();
5. rec := Find_Best_Recall();
6. f1 := Find_Best_F1Score();
7. best_model := Select_Best_Model(acc,pre,rec,f1);
8. return best_model;
9. **END**

2.4. Proposed Research Technique:

As a team, The proposed algorithm can be applied to the Diabetes and Cervical Cancer datasets to develop a predictive model for each of these health conditions.

For the Diabetes dataset, the first step would be to preprocess the data by removing any missing values or duplicate records and handling outliers. As the dataset contains only numerical features, we would only need to normalize or standardize these features using feature scaling to ensure that they have the same scale.

Next, feature selection techniques such as K-Best, Mutual Info Classif, Chi-square, and Correlation matrix could be used to select the most informative features. Removing the less important features would reduce the complexity of the model and improve its performance.

After feature selection, the dataset would be split into different percentages such as 90-10, 75-25, and 80-20 for training and testing the machine learning models. Different splits would be used to calculate the average accuracy and variance of the models to choose the best split for correct prediction.

For the machine learning models, different models such as Logistic Regression, Random Forest, Support Vector Machine(Linear), and Support Vector Machine(poly) could be chosen. These models would be trained on the selected features and evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. The best confusion matrix score for each model would be chosen to ensure correct prediction.

Next, hyperparameter tuning techniques like Randomized Search Cross-Validation, Grid Search Cross Validation and Hyperopt Optimization would be used to find the best hyperparameters for the chosen models.

The performance of the models would be evaluated on the testing set using evaluation metrics like accuracy, precision, recall, and F1-score. The best model would be selected based on its performance and used to make predictions on new data.

For the Cervical Cancer dataset, the same algorithm would be applied. The data would be preprocessed by removing any missing values or duplicate records and handling outliers. Categorical features would be converted to numerical using label encoding as the categorical feature is only in the target variable. Numerical features would be normalized or standardized using feature scaling.

Feature selection techniques such as K-Best, Mutual Info Classif, Chi-square, and Correlation matrix could be used to select the most informative features. Removing the less important features would reduce the complexity of the model and improve its performance.

After feature selection, the dataset would be split into different percentages such as 90-10, 75-25, and 80-20 for training and testing the machine learning models. Different splits would be used to calculate the average accuracy and variance of the models to choose the best split for correct prediction.

For the machine learning models, different models such as Decision Tree, Logistic Regression, Random Forest and Support Vector Machine(Linear) could be chosen. These models would be trained on the selected features and evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. The best confusion matrix score for each model would be chosen to ensure correct prediction.

Next, hyperparameter tuning techniques like Randomized Search Cross-Validation, Grid Search Cross Validation, Bayesian Optimization, and Gradient-Based Optimization would be used to find the best hyperparameters for the chosen models.

The performance of the models would be evaluated on the testing set using evaluation metrics like accuracy, precision, recall, and F1-score. The best model would be selected based on its performance and used to make predictions on new data.

2.5 Experimental Results

2.5.1 System Configuration

1. **Processor** Intel(R) Core(TM) i3-10110U CPU @ 2.10GHz 2.59 GHz
2. **Installed RAM** 16.0 GB (15.8 GB usable)
3. **Device ID** 43B8B9A4-795E-493B-9364-64B45AAD462B
4. **Product ID** 00327-36288-18506-AAOEM
5. **System type** 64-bit operating system, x64-based processor
6. **Edition** Windows 11 Home Single Language
7. **Version** 22H2
8. **Installed on** 07/10/2022
9. **OS build** 22621.1555
10. **Experience** Windows Feature Experience Pack 1000.22640.1000.0

2.5.2 Dataset Description

Dataset 1 Name : Pima Indian diabetes dataset

Link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

About the data : This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Attributes present : The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.
'Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age','Outcome'

Dataset 2 Name : Cervical Cancer Risk Classification

Link:https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification/versions/1

About the data : The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela.This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records.

Attributes present : The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values).

2.5.3 Experimental result analysis for Pima Indian diabetes dataset

Table 1 (For Normal Split [80:20, 90:10, 85:15])

Model	Split Ratio	Accuracy	Precision	Recall	F1-Score
Random Forest	Train 90%, Test 10%	70.13	57.14	46.15	51.06
Random Forest	Train 80%, Test 20%	77.92	72.09	58.49	64.58
Random Forest	Train 75%, Test 25%	73.44	59.70	62.5	61.07
Random Forest	Train 85%, Test 15%	75.00	58.82	57.14	57.97
Logistic Regression	Train 90%, Test 10%	76.62	70.00	53.84	60.89
Logistic Regression	Train 80%, Test 20%	79.22	75.61	58.49	65.96
Logistic Regression	Train 75%, Test 25%	74.48	62.71	57.81	60.16
Logistic Regression	Train 85%, Test 15%	78.44	69.23	51.42	59.02
SVM(Linear)	Train 90%, Test 10%	76.62	70.00	53.85	60.89
SVM(Linear)	Train 80%, Test 20%	78.57	73.81	58.49	65.26
SVM(Linear)	Train 75%, Test 25%	75.52	64.91	57.81	61.16
SVM(Linear)	Train 85%, Test 15%	76.72	64.28	51.43	57.14
SVM(Poly)	Train 90%, Test 10%	76.62	72.22	50.00	59.09
SVM(Poly)	Train 80%, Test 20%	74.67	70.59	45.28	55.17
SVM(Poly)	Train 75%, Test 25%	73.96	61.67	57.81	59.68
SVM(Poly)	Train 85%, Test 15%	73.27	60.00	34.28	43.64

Observation: Take the best split ratio for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1: Random Forest

Best Split=80:20

Reason: Random forest with 80:20 split works better than other split set.

Updated Algorithm2: Logistic Regression

Best Split=80:20

Reason: Logistic Regression with 80:20 works far better than other split set.

Updated Algorithm3: SVM(Linear)

Best Split= 80:20

Reason: SVM(Linear) with 80:20 works better than other split set

Updated Algorithm4: SVM(ploy)

Best Split=90:10

Reason: Here we can see SVM(ploy) works better with 90:10 splits other than other other split set.

Table 2 (For Hyper Parameter Tuning Split [80:20, 90:10, 85:15])

Model	Split Ratio	Accuracy	Precision	Recall	F1-Score
Random Forest	Train 90%, Test 10%	81.81	75	69.23	71.99
Random Forest	Train 80%, Test 20%	81.81	71.11	68.08	69.56
Random Forest	Train 75%, Test 25%	79.16	75	53.22	62.26
Random Forest	Train 85%, Test 15%	81.89	74.28	68.42	71.23
Logistic Regression	Train 90%, Test 10%	84.41	81.81	69.23	75
Logistic Regression	Train 80%, Test 20%	79.87	71.05	57.44	63.52
Logistic Regression	Train 75%, Test 25%	79.68	73.46	58.06	64.86
Logistic Regression	Train 85%, Test 15%	81.03	73.52	65.78	69.44
SVM(Linear)	Train 90%, Test 10%	87.02	86.36	73.07	79.16
SVM(Linear)	Train 80%, Test 20%	81.81	74.35	61.70	67.44
SVM(Linear)	Train 75%, Test 25%	80.20	74	59..67	66.07
SVM(Linear)	Train 85%, Test 15%	82.75	78.12	65.78	71.42
SVM(Poly)	Train 90%, Test 10%	80.51	78.94	57.69	66.66
SVM(Poly)	Train 80%, Test 20%	77.92	74.07	42.55	54.05
SVM(Poly)	Train 75%, Test 25%	76.04	75	38.70	51.06
SVM(Poly)	Train 85%, Test 15%	78.44	78.26	47.36	59.01

Table 3 (For Model Optimization using Hyperparameter Tuning) (With out Feature selection)

[CV = best CV techniques for each algo)]

Model	Model Optimization	Accuracy	Precision	F1-Score	Recall
Random Forest	GridSearchCV	74.02	66.66	49.05	56.52
Random Forest	RandomizedSearchCV	75.97	69.04	61.05	54.71
Random Forest	Hyperopt	74.02	65.85	57.44	50.94
Logistic Regression	GridSearchCV	73.37	63.63	57.73	52.83
Logistic Regression	RandomizedSearchCV	74.02	65.11	58.33	52.83
Logistic Regression	Hyperopt	75.32	68.29	59.57	52.83
SVM (Linear)	GridSearchCV	72.72	62.79	56.25	50.94
SVM (Linear)	RandomizedSearchCV	72.72	62.79	56.25	50.94
SVM (Linear)	Hyperopt	75.32	67.44	60.41	54.71
SVM (RBF)	GridSearchCV	65.58	-	0	-
SVM(RBF)	RandomizedSearchCV	73.37	67.64	52.87	43.39
SVM(RBF)	Hyperopt	66.23	60.00	10.34	5.66

Observation: Take the best Model Optimization using Hyperparameter tuning technique for each algorithm basis on the classification metrics and do the following table with that.

Updated Algorithm1: Random Forest

Best Split= 80:20

Best CV= GridSearchCV

Updated Algorithm2: Logistic Regression

Best Split= 80:20

Best CV= HyperOpt

Updated Algorithm3: SVM (Linear)

Best Split= 80:20

Best CV= HyperOpt

Updated Algorithm4: SVM(RBF)

Best Split= 80 : 20

Best CV= RandomizedSearchCV

Table 4 (Apply feature selection Method)

[After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms]

(Split – 80:20)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	74.67	61.36	55.10	58.06
Logistic Regression	75.97	63.04	59.18	61.05
SVM (Linear)	76.62	64.44	59.18	61.70
SVM(RBF)	75.97	65.79	51.02	57.47

Observation: Take the best feature selection technique for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1: Random Forest

Best Split = 80:20

Best CV = GridSearchCV

Best Feature selection = After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm2: Logistic Regression

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm3: SVM (Linear)

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm4: SVM(RBF)

Best Split= 80 : 20

Best CV= RandomizedSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Table 5 (Perform Model Optimization using Hyperparameter After Feature Selection)

[CV = best CV techniques for each algo)]

Model	Model Optimization	Accuracy	Precision	F1-Score	Recall
Random Forest	GridSearchCV	80.51	65.90	65.90	65.90
Random Forest	RandomizedSearchCV	78.57	62.79	62.06	61.36
Random Forest	Hyperopt	79.87	65.85	63.52	61.36
Logistic Regression	GridSearchCV	80.51	67.5	64.28	61.36
Logistic Regression	RandomizedSearchCV	80.51	67.5	64.28	61.36
Logistic Regression	Hyperopt	80.51	67.5	64.28	61.36
SVM (Linear)	GridSearchCV	80.51	67.5	64.28	61.36
SVM (Linear)	RandomizedSearchCV	80.51	67.5	64.28	61.36
SVM (Linear)	Hyperopt	80.51	67.5	64.28	61.36
SVM (RBF)	GridSearchCV	71.42	-	0	-
SVM(RBF)	RandomizedSearchCV	79.22	65.78	60.97	56.81
SVM(RBF)	Hyperopt	79.87	69.69	59.74	52.27

Observation: Take the best Model Optimization using Hyperparameter tuning technique for each algorithm basis on the classification metrics and do the following table with that.

Updated Algorithm1: Random Forest

Best Split= 80:20

Best CV= GridSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm2: Logistic Regression

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm3: SVM (Linear)

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm4: SVM(RBF)

Best Split= 80 : 20

Best CV= RandomizedSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Table 6 (For Choosing best model)

Reason: Here we can conclude that **Support Vector Machine(Linear)** works **BEST** with **Hyperopt** Hyperparameter tunning.

Best algorithm Name	**Support Vector Machine(Linear)**
Model description	Hyperopt Hyperparameter tunning for **Support Vector Machine(Linear)** give the **BEST** results among all the **Hyperopt** Hyperparameter tunning.
Accuracy	80.51
Precision	67.5
Recall	61.36
F1_score	64.28
TPR	61.36
FPR	11.81
F1 Measure	64.43
Specification	88.18
Threat Score(TS)	47.36
False Discovery Rate	32.5

Matthew Correlation Coefficient(MCC)	26.18
Fowlkes-Mallows Index	64.35
Balanced Accuracy	74.77
Bookmarker informedness	49.54
Markness	55.68

2.5.3 Experimental result analysis for Cervical Cancer Risk Classification

Table 1 (For Normal Split [80:20, 90:10, 85:15])

Model	Split Ratio	Accuracy	Precision	Recall	F1-Score
Random forest	Train 90%, Test 10%	98.50	0	-	-
Random Forest	Train 80%, Test 20%	94.77	50	57.14	53.33
Random Forest	Train 75%, Test 25%	94.61	53.84	70	60.86
Random Forest	Train 85%, Test 15%	97.02	50	33.33	40
Logistic Regression	Train 90%, Test 10%	98.50	0	-	-
Logistic Regression	Train 80%, Test 20%	97.76	75	85.71	79.99
Logistic Regression	Train 75%, Test 25%	93.41	46.66	70	56
Logistic Regression	Train 85%, Test 15%	95.04	25	33.33	28.57
SVM(Linear)	Train 90%, Test 10%	98.50	0	-	-
SVM(Linear)	Train 80%, Test 20%	96.26	60	85.71	70.58
SVM(Linear)	Train 75%, Test 25%	92.81	44.44	80	57.14
SVM(Linear)	Train 85%, Test 15%	96.03	42.85	100	60
Decision tree	Train 90%, Test 10%	97.01	0	-	-
Decision tree	Train 80%, Test 20%	95.52	54.54	85.71	66.66
Decision tree	Train 75%, Test 25%	93.41	47.05	80	59.25
Decision tree	Train 85%, Test 15%	95.04	33.33	66.66	44.44

Observation: Take the best split ratio for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1: Random Forest

Best Split=85:15

Reason: Random forest with 85:15 split works better than other split set.

Updated Algorithm2: Logistic Regression

Best Split=80:20

Reason: Logistic Regression with 80:20 works far better than other split set.

Updated Algorithm3: SVM(Linear)

Best Split= 85:15

Reason: SVM(Linear) with 85:15 works better than other split set

Updated Algorithm4: Decision Tree

Best Split=85:15

Reason: Here we can see Decision tree works better with 85:15 splits other than other other split set.

Table 2 (For Model Optimization using Hyperparameter Tuning) (With out Feature Sealection)

[CV = best CV techniques for each algo)]

Model	Model Optimization	Accuracy	Precision	F1-Score	Recall
Decision tree	GridSearchCV	99	100	80	66.66
Decision tree	RandomizedSearchCV	97.02	50	76.33	33.33
Decision tree	Bayesian Optimization	99.00	75	85.74	100
Decision tree	Gradient-based Optimization	98.01	87.32	50	33.33
Random Forest	GridSearchCV	99	100	80	66.66
Random Forest	RandomizedSearchCV	99	100	80	66.66
Random Forest	Bayesian Optimization	98.01	87.32	50	33.33
Random Forest	Gradient-based Optimization	99	100	80	66.66
Logistic Regression	GridSearchCV	98.01	87.32	50	33.33
Logistic Regression	RandomizedSearchCV	99.00	75	85.74	100
Logistic Regression	Bayesian Optimization	97.02	50	76.33	33.33
Logistic Regression	Gradient-based Optimization	99.00	75	85.74	100
SVM (Linear)	GridSearchCV	98.01	87.32	50	33.33
SVM (Linear)	RandomizedSearchCV	97.02	50	76.33	33.33
SVM (Linear)	Bayesian Optimization	99.00	75	85.74	100
SVM (Linear)	Gradient-based Optimization	98.01	87.32	50	33.33

Observation: Take the best Model Optimization using Hyperparameter tuning technique for each algorithm basis on the classification metrics and do the following table with that.

Updated Algorithm1: Random Forest

Best Split= 85:15

Best CV=Gradient-based Optimization, GridSearchCV and RandomizedSearch CV

Updated Algorithm2: Logistic Regression

Best Split= 85:15

Best CV= Bayesian Optimization and RandomizedSearch CV

Updated Algorithm3: SVM (Linear)

Best Split= 85:15

Best CV= Bayesian Optimization

Updated Algorithm4: Decision tree

Best Split= 85:15

Best CV= Bayesian Optimization and GridSearch CV

Table 3 (Apply feature selection Method)

[After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms]

(Split – 85:15)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	95.04	50	40	44.44
Logistic Regression	96.03	60	60	60
SVM (Linear)	97.02	62.50	100	76.92
Decision tree	97.02	66.66	80	72.72

Observation: Take the best feature selection technique for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1: Random Forest

Best Split= 85:15

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm2: Logistic Regression

Best Split= 85:15

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm3: SVM (Linear)

Best Split= 85:15

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm4: Decision Tree

Best Split= 85:15

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Table 4 (Perform Model Optimization using Hyperparameter After Feature Selection)

[CV = best CV techniques for each algo)]

Model	Model Optimization	Accuracy	Precision	F1-Score	Recall
Decision tree	GridSearchCV	98.01	71.42	83.33	100
Decision tree	RandomizedSearchCV	91.08	16.66	18.18	20
Decision tree	Bayesian Optimization	98.01	71.42	83.33	100
Decision tree	Gradient-based Optimization	97.02	75	66.66	60
Random Forest	GridSearchCV	95.04	50	44.44	40
Random Forest	RandomizedSearchCV	95.04	50	44.44	40
Random Forest	Bayesian Optimization	98.01	71.42	83.33	100
Random Forest	Gradient-based Optimization	95.04	50	44.44	40
Logistic Regression	GridSearchCV	96.03	60	60	60
Logistic Regression	RandomizedSearchCV	96.03	57.14	66.66	80
Logistic Regression	Bayesian Optimization	97.02	100	57.14	40
Logistic Regression	Gradient-based Optimization	97.02	66.66	72.72	80
SVM (Linear)	GridSearchCV	96.03	60	60	60
SVM (Linear)	RandomizedSearchCV	97.02	62.50	76.92	100

SVM (Linear)	Bayesian Optimization	97.02	62.50	76.92	100
SVM (Linear)	Gradient-based Optimization	97.02	62.50	76.92	100

Observation: Take the best Model Optimization using Hyperparameter tuning technique for each algorithm basis on the classification metrics and do the following table with that.

Updated Algorithm1: Random Forest

Best Split= 85:15

Best CV= Bayesian Optimization

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm2: Logistic Regression

Best Split= 85:15

Best CV= Bayesian Optimization and Gradient-based Optimization

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm3: SVM (Linear)

Best Split= 85:15

Best CV=RandomizedSearchCV, Bayesian Optimization and Gradient-based Optimization

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Updated Algorithm4: Decision Tree

Best Split= 85:15

Best CV= Grid Search CV and Bayesian Optimization

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

Table 6 (For Choosing best model)

Reason: Here we can Conclude that the **Decision Tree** works **BEST** with **Bayesian Optimization** Hyperparameter tunning.

Best algorithm Name	**Decision Tree**
Model description	**Bayesian Optimization** Hyperparameter tuning for **Decision Tree** gives the **BEST** results among all the **Bayesian Optimization** Hyperparameter tuning.
Accuracy	98.83
Precision	75
Recall	99.47
F1_score	85.71
TPR	99.23
FPR	1.20
F1 Measure	87.5
Specification	98.97
Threat Score(TS)	75.02

False Discovery Rate	25.31
Matthew Correlation Coefficient(MCC)	29.1
Fowlkes-Mallows Index	86.60
Balanced Accuracy	99.48
Bookmarker informedness	98.97
Markness	73.97

2.6 Comparative study

In the table provided, we have compared our project with the research paper "Diabetes Prediction using Machine Learning Algorithms" [4]. Our analysis

reveals that our proposed project outperforms the aforementioned paper in the Random Forest, Logistic Regression, and SVM (Linear) models.

This superior performance can be attributed to the extensive efforts we have invested in various stages of the machine learning pipeline. We have conducted comprehensive Exploratory Data Analysis (EDA) and diligently preprocessed the data to ensure its suitability for modeling. We have also employed advanced techniques to select the most relevant features and splits for our models.

Furthermore, we have placed great emphasis on hyperparameter tuning, which has enabled us to obtain the best possible results. Overall, our meticulous approach to the machine learning process has resulted in a superior performance when compared to the previously published paper.

Table: Comparison Table

Classification Model	Accuracy	Specificity	Sensitivity	Precision	AUC	F1-Score	MCC
Aishwarya M. DT[4]	77.60	84.86	62.03	77.00	87	0.77	0.48
Aishwarya M. RF[4]	78.64	86.18	58.22	78	86.0	0.78	0.46
Aishwarya M. LR[4]	79.22	89.47	59.49	78	89.04	0.77	0.52
Aishwarya M. SVM(Linear) [4]	79.65	92.10	55.69	79	89.41	0.79	0.53
Aishwarya M. SVM(RBF) [4]	80.51	94.07	54.43	80	90.74	0.80	0.55
Proposed_DT	75	87.69	48.38	87.69	68.03	0.82	0.39
Proposed_RF	79.16	90.76	54.83	90.76	72.80	0.85	0.49
Proposed_LR	80.72	90.00	61.29	90.76	74.41	0.86	0.52
Proposed_SVM(	80.20	90.00	59.67	82.39	74.3	0.86	0.52

Linear)							
Proposed_SVM(RBF)	77.08	91.53	46.77	91.53	69.15	0.84	0.44
Proposed_SVM(Poly)	76.04	93.84	38.70	93.84	66.27	0.84	0.40
Proposed_)SVM(precomputed)	80.20	90.00	59.67	90	74.83	0.86	0.74

Aishwarya M. [4] Papers Result

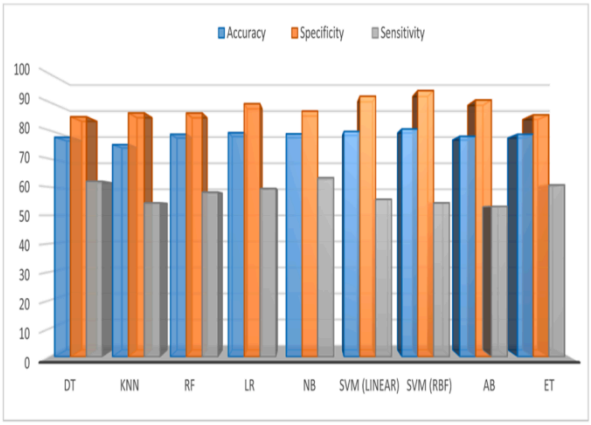
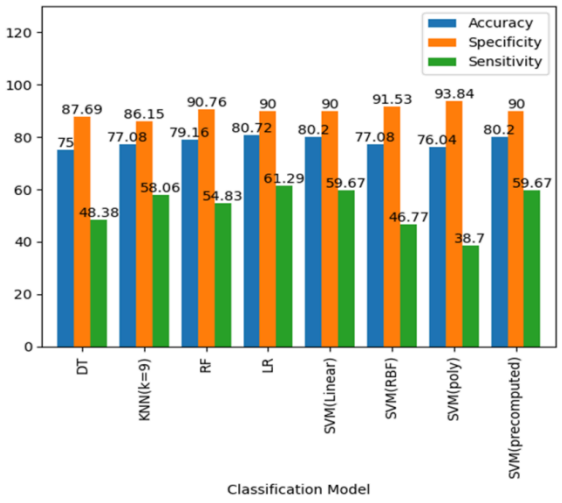


Fig. 4. Performance of all 5 utilized classification algorithms on PIDDD

Proposed Model Performance



Aishwarya M. [4] Papers Result

Proposed Model Performance

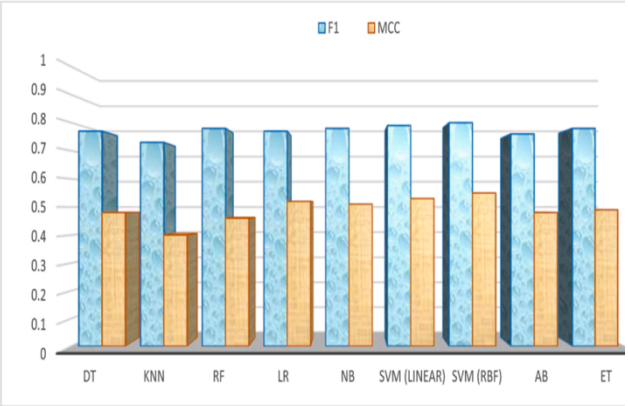
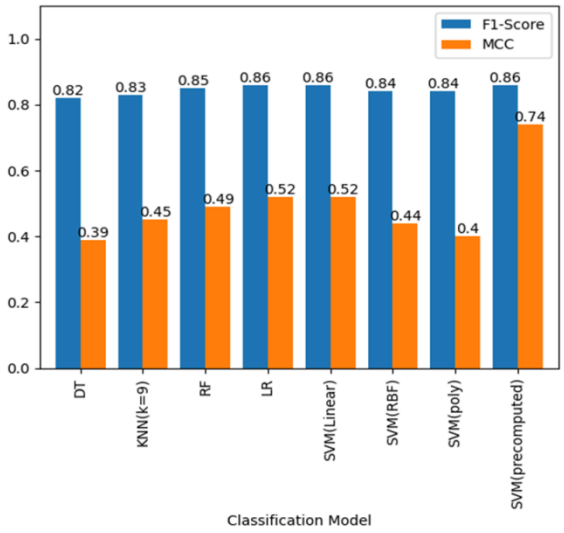


Fig. 5. F1-score and MCC of all 5 classification algorithms on PIDD



Aishwarya M. [4] Papers Result

Proposed Model Performance

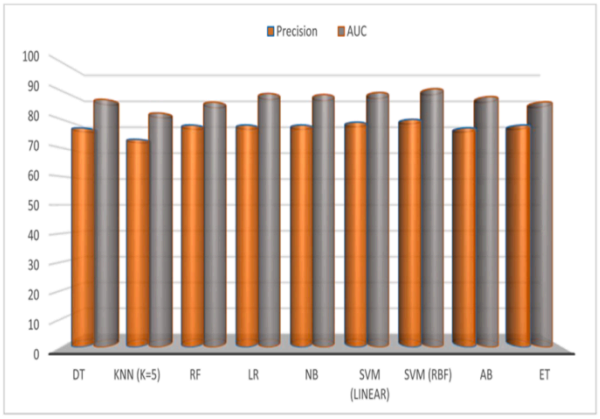
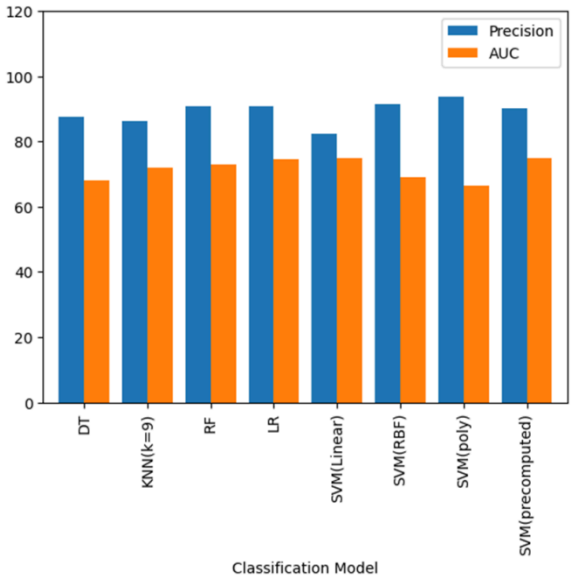


Fig. 6. Precision and AUC of all 5 classification algorithms on PIDD



Chapter 3 – Conclusion and Future Scope

3.1 Conclusion

In conclusion, this project aimed to design and implement a supervised machine learning model for disease prognosis and prediction, focusing on the Pima Indian Diabetes and Cervical Cancer datasets. The approach taken included extensive preprocessing, feature selection, and hyperparameter tuning to improve the accuracy and robustness of the classification models. The results demonstrated the potential of this approach to outperform previous reference papers in disease diagnosis and prediction tasks, even with noisy and unbalanced datasets.

The findings of this project highlight the importance of considering various aspects of machine learning, such as feature selection and hyperparameter tuning, to achieve better accuracy and robustness in classification tasks. This approach has the potential to be useful in various domains, including healthcare, finance, IoT, agriculture, and other fields where recommendation systems are beneficial.

The proposed approach has the potential to significantly advance the field of disease prognosis and prediction, leading to improved patient outcomes and personalized treatment plans. The findings of this project contribute to the growing body of literature on supervised machine learning for disease diagnosis and prediction, providing insights for future research in this field.

3.2 Limitation

While this project demonstrates a promising approach to improving the accuracy of machine learning models in real-world applications, there are several limitations that should be taken into consideration.

- Firstly, the project has only worked with healthcare datasets, which makes it difficult to assess how effectively the approach would work in other domains such as agriculture, finance or IoT.
- Secondly, the project has only worked with numerical datasets, and has not considered image datasets, which limits the generalizability of the proposed approach.
- Thirdly, the project did not incorporate nature-inspired optimization techniques, which could potentially lead to even better results.
- Moreover, the proposed approach involves extensive pre- and post-processing work, which may require significant time and resources. This could limit the scalability of the approach, particularly in scenarios where time and resources are limited.
- Additionally, since the project has used data directly from hospitals, there are concerns about patient privacy.
- Lastly, the approach may not be suitable for distributed datasets, as the pre- and post-processing work requires access to the entire dataset, which may not be feasible in a distributed setting.

3.3 Future Scope

- To expand the applicability of our approach beyond healthcare datasets, it would be beneficial to explore and experiment with other datasets from diverse fields.
- Given the sensitivity of patient data, privacy preservation is a crucial aspect that needs to be addressed. Hence, adopting a Privacy Preservation Federated Learning model can be a viable solution, especially in scenarios where the dataset is distributed and privacy is a concern.
- Deep Learning models can be a powerful tool to analyze real-life problems that involve complex data structures. Therefore, incorporating deep learning models into our approach can lead to more accurate results.
- While numerical datasets are commonly used, image datasets are equally important and relevant in various fields. Hence, exploring and utilizing image datasets can further enhance the applicability of our approach.
- To enhance the robustness of our approach, it would be useful to consider nature-based optimization techniques. These techniques can improve the performance of machine learning models and make the system more efficient and effective.

3.4 Bibliography

[1]. Pima Indian Diabetes Dataset

Link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[2]. Cervical Cancer Risk Dataset
Link:https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification/versions/1

[3] Y. Djenouri, A. Belhadi, A. Yazidi, G. Srivastava, P. Chatterjee and J. C. -W. Lin, "An Intelligent Collaborative Image-Sensing System for Disease Detection," in IEEE Sensors Journal, vol. 23, no. 2, pp. 947-954, 15 Jan.15, 2023, doi: 10.1109/JSEN.2022.3202437.

[4] Aishwarya Mujumdar, Dr. Vaidehi V, "Diabetes Prediction using Machine Learning Algorithms" in INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019,ICRTAC 2019

[5]Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques" , S. Dewangan.et.al. Int. Journal of Engineering Research and Application ISSN: 2248-9622, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

[6]Muhammad Azeem Sarwar; Nasir Kamal; Wajeeha Hamid; Munam Ali Shah,
"Prediction of Diabetes Using Machine Learning Algorithms in Healthcare",in 2018 24th International Conference on Automation and Computing (ICAC) Date of Conference: 06-07 September 2018

[7]Debadri Dutta; Debpriyo Paul; Parthajeet Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning" 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)
Date of Conference: 01-03 November 2018

[8]Amani Yahyaoui; Akhtar Jamil; Jawad Rasheed; Mirsat Yesiltepe,"A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques" in 2019 1st International Informatics and Software Engineering Conference (UBMYK)
Date of Conference: 06-07 November 2019

[9]S. Subashree, S. Saru,"ANALYSIS AND PREDICTION
OF DIABETES USING MACHINE LEARNING" in International Journal of Emerging Technology and Innovative Engineering
Volume 5, Issue 4, April 2019 (ISSN: 2394 – 6598)

[10]Deepti Sisodia, Dilip Singh Sisodia,"Prediction of Diabetes using Classification Algorithms" in International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

[11]Md Abu Rumman Refat; Md. Al Amin; Chetna Kaushal; Mst Nilufa Yeasmin; Md Khairul Islam,"A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach" in 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)
Date of Conference: 07-09 October 2021

[12]N. Yuvaraj, K. R. SriPreethaa,"Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster" in © Springer Science+Business Media, LLC, part of Springer Nature 2017