Leveraging Machine Learning to Predict Disease Risk Levels from Multiple Eye Condition Indicators

Mudumala Varnika Narayani
Dept. of Computer Science and Engineering,
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22035@bl.students.amrita.edu

Nunnaguppala Rohit
Dept. of Computer Science and Engineering,
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22040@bl.students.amrita.edu

Naga Ruthvika Durupudi
Dept. of Computer Science and Engineering,
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22036@bl.students.amrita.edu

Tejashwini Vadeghar
Dept. of Computer Science and Engineering,
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
tejashwini.vadeghar@gmail.com

Jyotsna C.
Dept. of Computer Science and Engineering,
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
c_jyotsna@blr.amrita.edu

Aiswariya Milan K.
Dept. of Computer Science and Engineering,
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
m_aiswariya@blr.amrita.edu

***Abstract*—This paper presents a review of machine learning methodologies applied to disease risk evaluation, leveraging ocular condition indices derived from the RFMiD dataset. To enhance the performance of classification, a formal and structured mechanism of statistical analysis and also the method of model prediction was applied. Data mining techniques which include Decision Trees, Random Forest, Naive Bayes, Gradient Boosting, and Stacking Classifiers were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. For the cases of coarse models applied to healthcare XAI approaches such as SHAP and LIME were used to gain better interpretability. This paper points to the potential of machine learning for disease risk modelling and provides recommendations for generating accurate and easily-interpreted models in order to increase the diagnostic accuracy and, therefore, the quality of care for patients.**

Keywords: **Disease classification, Retinal Fundus MultiDisease Image Dataset, Centroid analysis, Minkowski distance, Explainabe AI, Diabetic Retinopathy**

I. Introduction

Machine learning (ML) is changing the way of analyzing and predicting complex patterns in data, making it a key tool in healthcare [1, 2]. However promising, choosing the optimal ML approach for specific tasks remains challenging, particularly in healthcare where accuracy and transparency are critical [3].

Predicting illness risk facilitates early detection and prevention, enabling timely intervention to minimize complications. Prioritizing high-risk patients allows for efficient resource allocation and personalized care. This study leverages machine learning models (Decision Trees, Random Forest, Gradient Boosting) on the RFMiD dataset to assess disease risks. SHAP and LIME enhance model reliability and interpretability, demonstrating machine learning's potential to improve diagnostic accuracy. The analysis focuses on key markers such as Macular Hole, Epiretinal Membrane, Retinitis Pigmentosa, and Choroidal Neovascularization, crucial for improved retinal disease diagnosis and treatment.

The Retinal Disease Classification dataset includes key markers for assessing the risk of diseases such as Age-Related Macular Degeneration (ARMD), Diabetic Retinopathy (DR), Hypertensive Retinopathy, Glaucoma, Central Serous Retinopathy (CSR), and Retinal Vein Occlusion (RVO). These conditions, respectively, involve central retinal degeneration, retinal blood vessel damage from diabetes, retinal effects of high blood pressure, optic nerve damage from increased eye pressure, fluid accumulation behind the retina, and blocked retinal veins affecting vision.

Optimizing ML performance, ensuring generalization, and guaranteeing interpretability, particularly in healthcare, remain key challenges. Addressing these challenges is crucial for realizing ML's real-world potential [4].

SHAP and LIME improve machine learning transparency by elucidating prediction formation, going beyond simple feature importance. SHAP quantifies each feature's contribution to a prediction, enabling global and local interpretability, unlike LIME which is limited to individual instances. In healthcare, SHAP fosters trust by providing clear rationales for critical predictions. This research explores complex ML models like decision trees, random forests, and gradient boosting for disease risk prediction, evaluating their performance using accuracy, precision, recall, F1, and ROC AUC. Naive Bayes demonstrates superior predictive ability and emerges as the top model [5].

This study employs a research pipeline encompassing preprocessing, feature engineering, and stratified k-fold cross-validation for robust disease risk prediction using advanced ML models on healthcare datasets. It highlights the importance of interpretability tools (SHAP and LIME) for transparent healthcare decisions and proposes methods to deploy ML models that balance performance with explainability, ultimately aiming to improve diagnostics, decision-making, and patient outcomes.

II. Literature Survey

Supervised machine learning has improved disease prediction in healthcare. While Support Vector Machines are common, Random Forests usually yield better accuracy. Future research should aim to resolve these issues for more trustworthy models [6]. Machine learning has improved, especially in predicting diseases like breast cancer and heart disease [7]. Random Forest achieves 88.5% accuracy in cardiovascular disease prediction with the Cleveland HD dataset. Future work should focus on improving feature selection for better model accuracy [8]. Random Forest achieved the best AUC of 0.96 in predicting Metabolic Syndrome in perimenopausal women. SHAP values highlighted key risk factors like waist circumference and fasting blood glucose. Future work should enhance feature importance for clearer predictions [9]. Unsupervised learning models, including DBSCAN and Bayesian Gaussian Mixture, show potential for unstructured healthcare data. Further work is necessary to apply these models to more complex and diverse healthcare datasets [10]. In liver disease prediction, SVM has proven to be the most accurate model. Further studies should allow for more and better algorithms for preg Processing as well as having better algorithms for precision [11]. Imbalanced classes were balanced with Ensemble methods and SMOTE with a high accuracy of the Voting classifier. of 80.1%. Future work could focus on diversifying the models and incorporating more real-world datasets to improve accuracy [12].

SVM had better accuracy than Logistic Regression and Decision Trees in the prediction of liver disease. They reckon that larger data sets should be used for improving versatility and that future studies should focus on improving the algorithms for different uses [13]. Challenges in healthcare data classification, including issues of accuracy, scalability, and interpretability, were addressed using a hybrid approach combining Random Forest and K-Means clustering. This method outperformed traditional models. Future research should validate this approach using precision-recall curves and real-world data [14]. Random Forest demonstrated an impressive AUC of 89.05% in predicting chronic diseases, offering promising results for imbalanced datasets. Future research should focus on exploring further ensemble methods to enhance chronic disease prediction [15]. Machine learning algorithms across various diseases were evaluated, with SVM being the most used and Random Forest achieving high accuracy. Future studies should focus on hyperparameter tuning and refining sub-classifications to improve model performance [16]. Ensemble learning techniques, including Bagging, Boosting, Stacking, and Voting, were reviewed for disease prediction, with Stacking showing the highest accuracy. Future research should aim to explore more efficient ensemble methods to enhance predictive performance.[17]. Ensemble methods such as Bagging, Boosting, Stacking, and Voting continue to be powerful tools for disease prediction, with Stacking standing out due to its accuracy. Future research should focus on improving the robustness and accuracy of these methods [18]. Challenges in machine learning for disease prediction include issues related to model interpretability, data quality, and generalizability. Future work should focus on improving model transparency and balancing datasets to ensure fairer predictions [19].

The review focuses on enhancing accuracy, interpretability, and robustness in healthcare applications, highlighting noteworthy developments in machine learning-based disease risk prediction. Even while algorithms like Random Forest and Support Vector Machines (SVM) continuously show great accuracy on a variety of datasets, problems including class imbalance, uneven data quality, and the requirement for explainable models still exist. While tools like SHAP and LIME increase model interpretability, recent work have used ensemble techniques like Stacking and Gradient Boosting to improve predictive performance. However, for practical application in the healthcare industry, striking a balance between explainability and accuracy is still crucial. In particular, when dealing with high-dimensional and unbalanced datasets in healthcare diagnostics, future research paths will focus on improving feature selection, correcting data biases, and optimizing models for wider generalizability.

III. Methodology

The systematic model development included data preprocessing (handling missing values, applying StandardScaler). Minkowski distance, scatter plots, and histograms were used to explore feature relationships. An 8:2 stratified train/test split was used. Decision Tree, Random Forest, Gradient Boosting, and Naive Bayes were evaluated using accuracy. Random Forest excelled. SHAP and LIME improved interpretability via feature contribution analysis. Grid Search optimized hyperparameters. ROC curves and SHAP plots visualized results. Data augmentation enhanced training. GridSearchCV and RandomizedSearchCV tuned models. Ensemble methods like Gradient Boosting and Random Forest increased accuracy.

A. Data Loading and Preprocessing

The ”Retinal Disease Classification” dataset comprises 1,920 records and 47 numerical features, including the target variable ”Disease_Risk.” Key features include ”DR,” ”ARMD,” and ”DN.” Feature standardization via StandardScaler, focusing on features correlated with ”Disease_Risk,” was applied to enhance predictive accuracy. The dataset contains no missing values, making it well-suited for machine learning.

B. Feature analysis

The dataset contains columns for an ID, a risk score (Disease_Risk), and binary indicators for diseases (e.g., DR, ARMD, MH).

1. Histogram Analysis: Histogram plots were made to visualize feature distributions like Diabetic Retinopathy, providing insights into dataset conditions. Mean and variance calculations assessed central tendency and variability.

2. Scatter Plots: Scatter plots were used to explore relationships between features, like DR and ARMD, showing potential correlations and clustering in the data.

3. Minkowski Distance Calculation: The Minkowski distance was computed for features (e.g., DR and ARMD) with r values from 1 to 10 to assess similarity and support feature selection.

C. Train-Test split

The dataset was meticulously processed to select relevant features and targets, while systematically removing any unnecessary or redundant columns that could introduce noise, ensuring a more efficient and accurate analysis. A stratified 80/20 train-test split was employed to preserve the original class distribution, thus mitigating any potential bias and enhancing the model's ability to generalize to new, unseen data. To ensure the consistency and reliability of the results, random seeds were set during the entire process, allowing for reproducibility across experiments. This well-structured and balanced split of the data enabled seamless execution of model training, precise hyperparameter tuning, and comprehensive classifier evaluation, all of which are clearly depicted in Figure 1.

D. Implementation of Models

1. Decision Trees: The data is partitioned using either entropy or Gini impurity as the splitting criterion. To optimize data splitting, GridSearchCV is used to tune parameters for entropy or Gini impurity-based splitting, ensuring optimal performance. Insect identification is evaluated using accuracy, precision (insect ID accuracy), recall (insect detection rate), F1-score (balancing precision/recall), ROC-AUC (class discrimination), feature significance (identifying key features), and a confusion matrix (visualizing TP, TN, FP, FN per class).

2. Random Forest Classifier: We implemented a bagged decision tree model, tuning hyperparameters like n_estimators and max_depth using GridSearchCV. Evaluation metrics included accuracy, precision, recall, F1-score, ROC-AUC, and feature importance.

3. Gradient Boosting Classifier: Iteratively grows trees, correcting errors from previous iterations (parameterized by GridSearchCV-optimized estimators and learning rate). Excels at capturing two- and three-way feature interactions relevant to disease risk, outperforming other models.

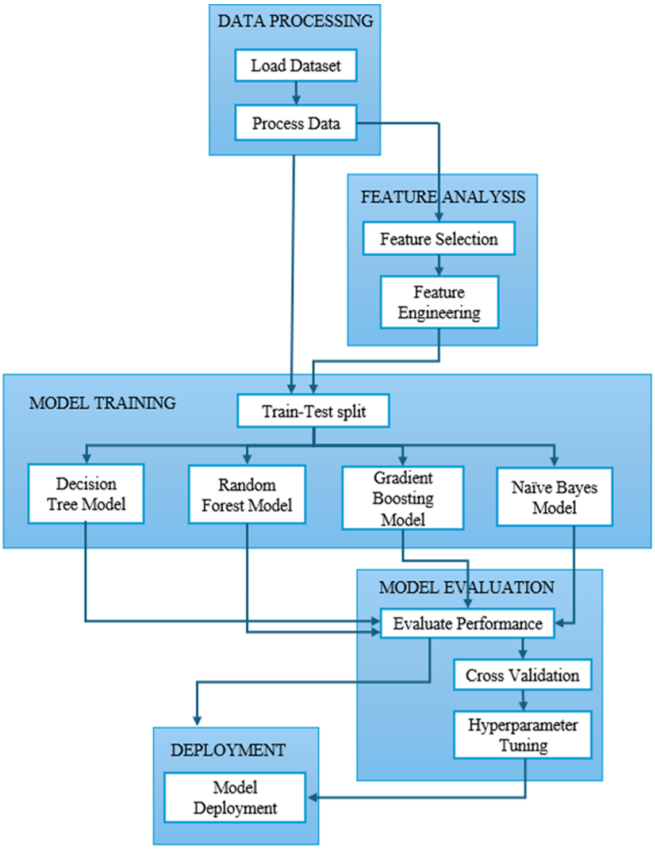

Fig. 1: Architecture diagram

Algorithm 1 Gradient Boosting Classifier Implementation

Dataset $D = \{(X, y)\}$, n_estimators, learning_rate, max_depth Trained Gradient Boosting model and performance metrics

Step 1: Load Dataset

Load the dataset D and define X and y as features and labels.

Step 2: Preprocess Data

Standardize the features X using a scaler to ensure uniformity in feature scaling.

Step 3: Initialize Gradient Boosting model

Define the Gradient Boosting model with the given $n\_estimators$, $learning\_rate$, and max_depth.

Step 4: Train the Model

Fit the Gradient Boosting model on the preprocessed training data.

Step 5: Make Predictions

Predict the labels on the test data using the trained Gradient Boosting model.

Step 6: Evaluate Performance

Evaluate the model's performance using accuracy, precision, recall, F1-score, and ROC-AUC.

Step 7: Visualize Results

Plot the ROC curve and confusion matrix to visualize model performance.

4. Naive Bayes Classifier: A simple probabilistic classifier based on the Naïve Bayes theorem which assumes that the features are independent,.bz Working well for text classification and large datasets, it is applied with smoothing to work appropriately with zero probabilities.

Algorithm 2 Naive Bayes Classifier Implementation

Dataset $D = \{(X, y)\}$, alpha Trained Naive Bayes model and performance metrics

Step 1: Load Dataset

Load the dataset D and define X and y as features and labels.

Step 2: Preprocess Data

Standardize the features X using a scaler if necessary.

Step 3: Initialize Naive Bayes model

Define the Naive Bayes model with the given α (smoothing parameter).

Step 4: Train the Model

Train the Naive Bayes model on the training data.

Step 5: Make Predictions

Use the trained Naive Bayes model to predict the labels on the test data.

Step 6: Evaluate Performance

Evaluate the model's performance using accuracy, precision, recall, and confusion matrix.

Step 7: Visualize Results

Plot a confusion matrix to visualize the performance of the model.

E. Model Evaluation and Comparison

Model performance was evaluated using multiple metrics:

- **Accuracy:** Measures the proportion of correct predictions.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall:** The ratio of correctly predicted positive observations to the actual positives.
- **F1-Score:** The weighted average of precision and recall, useful for imbalanced datasets.
- **ROC-AUC:** Measures the ability of the model to distinguish between classes.

F. Model Interpretation and Explainability

SHAP and LIME improve model interpretability by clarifying predictions and feature importance, aiding healthcare professionals in understanding patient risk. Traditional models like Random Forest and Gradient Boosting often lack clarity. This study applies SHAP and LIME to show how features like Age-Related Macular Degeneration impact outcomes, revealing how high DR values increase disease risk. SHAP plots also highlight key risk factors, ensuring models are both accurate and interpretable, which is vital in healthcare.

G. Hyperparameter Tuning

Parameters such as 'n_estimators' for random forest, 'max_depth'for Decision Trees, and learning rate for the Gradient Boosting model were optimized using GridSearchCV and RandomizedSearchCV.

H. Deployment and Visualization

A web interface to the disease risk predictions was created with explanation tools allow such as confusion matrix and ROC curve for analysis.

IV. Result analysis

This section analyzes the machine learning models used, including feature analysis, performance evaluation, and interpretability via SHAP and LIME, highlighting key observations.

A. Feature Analysis

The initial step in analyzing the dataset involved examining features through histograms to assess the frequency of Diabetic Retinopathy (DR) indicators. Mean and variance calculations highlighted central tendencies and variability. Figure 2 shows DR presence, highlighting features like retinal lesions and vascular changes.

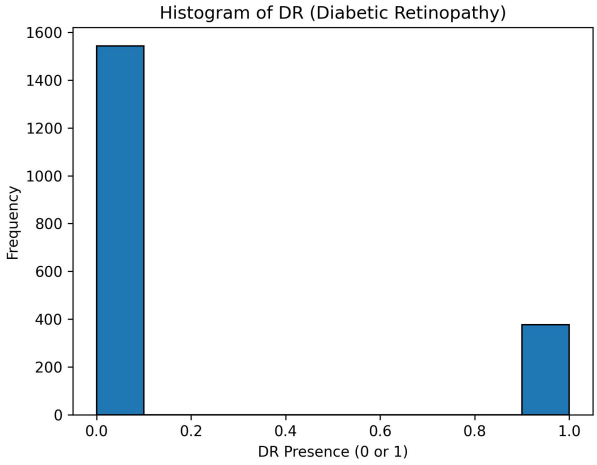

Fig. 2: Presence of Diabetic Retinopathy (DR) and Key Features

Minkowski distance calculations for r values ranging from 1 to 10 were performed to quantify the similarity between features. This helped in identifying closely related attributes for effective feature selection. Figure 3 visualizes the Minkowski distance metrics between Diabetic Retinopathy (DR) and Age-Related Macular Degeneration (ARMD), illustrating the potential to differentiate between conditions based on feature similarity.

Scatter plots were also generated to identify potential correlations between disease risk factors and the presence of Diabetic Retinopathy (DR). Figure 4 shows a scatter plot highlighting the relationship between these factors, providing insights into possible clinical connections.

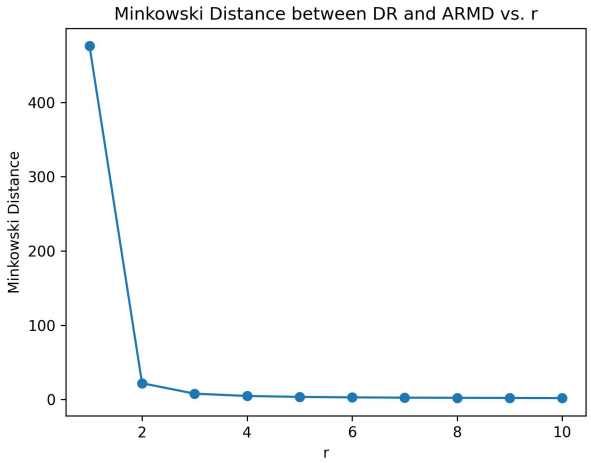

Fig. 3: Minkowski Distance Metric Between DR and ARMD

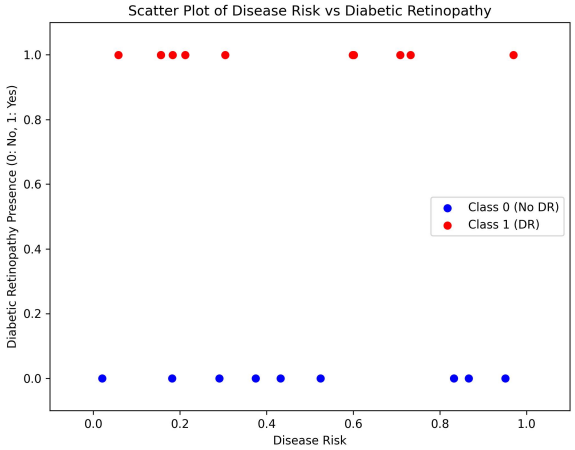

Fig. 4: Scatter Plot of Disease Risk Factors vs. DR

B. Models Applied

Several machine learning models were applied, ranging from traditional classifiers to more complex ensemble methods. Each model was assessed using various metrics, such as accuracy, precision, recall, and F1-score.

1) Decision Trees: Decision Trees were tested on real-world datasets. To avoid overfitting, constraints like maximum depth were imposed, promoting better generalization. Figure 5 depicts a simplified tree structure that retains predictive power. The entropy criterion was also used for node splitting, measuring information gain to identify the best attribute for partitioning.This method enhances the model's decision-making process, as demonstrated in Figure 6.

C. Model Performance

The Table 1 below summarizes the performance metrics for each model, highlighting the effectiveness of each approach in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

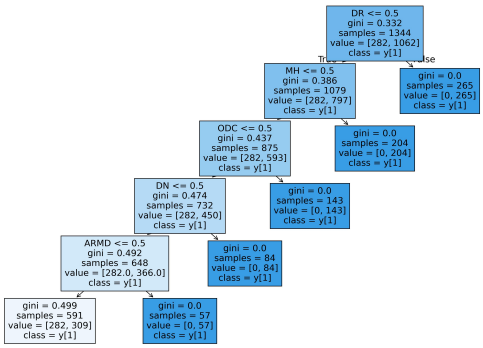

Fig. 5: Decision Tree with Maximum Depth Constraint

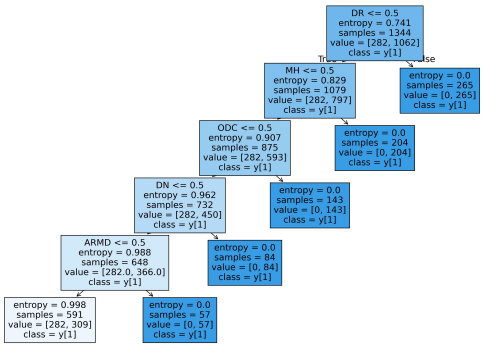

Fig. 6: Decision Tree with Entropy Criterion for Node Splitting

TABLE I: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	98.7%	1.0000	98.3	99.1%	0.9919
Random Forest	98.4%	1.0000	98	99%	0.9996
Naive Bayes	99.7%	1.0000	99.6	99.8%	0.9982
Gradient Boosting	97.4%	1.0000	96.7	98.3%	0.9884

D. Key Observations

Naive Bayes was especially helpful for categorizing illness risk because of its high accuracy, recall, and ROC-AUC. SHAP provided a global view of feature importance, highlighting the most impactful features, while LIME offered localized explanations, clarifying individual predictions by analyzing feature contributions.

E. Model Interpretability

1) SHAP Analysis: SHAP values assess each feature's role in model predictions, illustrating decision-making insights. By using Shapley values, SHAP measures the impact of features. SHAP plots highlight influential features and their contributions, enhancing transparency and identifying key risks.

2) LIME Explanation: LIME (Local Interpretable Model-Agnostic Explanations) improves understanding of complex models by producing local explanations for predictions. It modifies input data and observes output changes to show feature impacts. A LIME plot visually depicts these explanations for specific cases, enhancing user trust with clear insights.

V. Conclusion

This study shows Naive Bayes is an effective and efficient model for predicting illness risk. While more complex ensemble methods are common, Naive Bayes stands out for its speed, interpretability, and performance, particularly with large datasets. Although it doesn't surpass ensemble models, its accuracy and high recall make it valuable in healthcare, especially for timely forecasts. Naive Bayes remains a strong contender in disease risk prediction, balancing computational efficiency with reliable outcomes.

Future studies could integrate CNNs to better detect subtle patterns in retinal images and use real-time data from wearables to enhance disease risk prediction accuracy.Additionally, combining real-time data from wearable devices could improve the accuracy and timeliness of disease risk prediction models.

References

[1] Vinayak, Vineet, Mohan Paliwal, J. Amudha, and C. Jyotsna. "Prediction of Neuro Cognitive Disorders using Supervised Comparative Machine Learning Model & Scanpath Representations." In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), pp. 1-5. IEEE, 2023.

[2] Nishitha, U., Revanth Kandimalla, and C. Jyotsna. "Automobile Price Prediction using Machine Learning with Data Visualization." In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-7. IEEE, 2023.

[3] Vinayak, Vineet, and C. Jyotsna. "Consumer Complaints Classification using Deep Learning & Word Embedding Models." In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2023

[4] Varshith, Tummala, Thanush S. Koneri, Uppalapati Dhanush, Sriramoju Rahul, and C. Jyotsna. "DeepPave: Detection of Potholes Using Deep Learning Techniques." In 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1441-1446. IEEE, 2024.

[5] Sohel, Mohammad Aman, Raghupatruni Sai Madhukar, Sai Muralidhar Batchu, and C. Jyotsna. "Unmasking Emotions: Deep Neural Networks for Image-Based Emotion Recognition." In 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1045-1051. IEEE, 2024.

[6] Uddin, Shahadat, et al. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19.1 (2019): 1-16.

[7] Saranya, G., and A. Pravin. "A comprehensive study on disease risk predictions in machine learning." International Journal of Electrical and Computer Engineering 10.4 (2020): 4217.

[8] Karthick, K., et al. "[Retracted] Implementation of a Heart Disease Risk Prediction Model Using Machine Learning." Computational and Mathematical Methods in Medicine 2022.1 (2022): 6517716.

[9] Xiaoxue, Wang, et al. "Risk prediction model of metabolic syndrome in perimenopausal women based on machine learning." International Journal of Medical Informatics 188 (2024): 105480.

[10] Lu, Haohui, and Shahadat Uddin. "Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets." Health and Technology 14.1 (2024): 141-154.

[11] Modhugu, Venugopal Reddy, and Sivakumar Ponnusamy. "Comparative Analysis of Machine Learning Algorithms for Liver Disease Prediction: SVM, Logistic Regression, and Decision Tree." Asian Journal of Research in Computer Science 17.6 (2024): 188-201.

[12] Dritsas, Elias, and Maria Trigka. "Supervised machine learning models for liver disease risk prediction." Computers 12.1 (2023): 19.

[13] Thakur, Anjali, et al. "A Hybrid Approach for Heart Disease Detection using K-Means and K-NN Algorithm." American Journal of Electronics & Communication 4.1 (2023): 14-21.

[14] Manikandan, P. "Medical big data classification using a combination of random forest classifier and k-means clustering." International Journal of Intelligent Systems and Applications 10.11 (2018): 11.

[15] Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. "Predicting disease risks from highly imbalanced data using random forest." BMC medical informatics and decision making 11 (2011): 1-13.

[16] Uddin, Shahadat, et al. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19.1 (2019): 1-16.

[17] Mahajan, Palak, et al. "Ensemble learning for disease prediction: A review." Healthcare. Vol. 11. No. 12. MDPI, 2023.

[18] Pudjihartono, Nicholas, et al. "A review of feature selection methods for machine learning-based disease risk prediction." Frontiers in Bioinformatics 2 (2022): 927312.

[19] https://www.kaggle.com/datasets/andrewmvd/retinal-disease-classification