Comparative Evaluation of Predictive Models on Kidney, Lung Cancer and Heart Disease

Prof. Pradnya Bhangale
Assistant Professor
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, Maharashtra, India
pyb@somaiya.edu

Aarya Shah
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, Maharashtra, India
aarya19@somaiya.edu

Dev Patel
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, Maharashtra, India
dev03@somaiya.edu

Parth Shah
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, Maharashtra, India
pms2@somaiya.edu

Sagar Salvi
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, Maharashtra, India
sagar.salvi@somaiya.edu

***Abstract*—This study supports advances in machine learning to improve early detection and treatment planning for lung cancer, cardiovascular disease, and kidney disease. We compare traditional models such as decision trees and logistic regression with complex techniques such as support vector machines, random forests, and KNN and evaluate them on publicly available data. This hybrid approach uses random forest and decision tree classifiers, leveraging adaptive learning to improve model accuracy. Results showed high prediction accuracy for kidney disease and lung cancer , while prediction accuracy for heart disease was average . This difference indicates the need for better work and more information. Future studies will focus on improving cardiovascular models, addressing data uncertainty, and integrating predictive models into clinical practice to support early diagnosis and personalized treatment to improve patient outcomes. This study demonstrates the potential for machine learning to have a major impact on diagnosis and patient management.**

Keywords—Machine learning, Lung cancer, Cardiovascular Disease, Kidney Disease, Prediction Accuracy, Hybrid Model

I. INTRODUCTION

Efficient classification is important in the medical field — especially in the case of conditions such as kidney disease, lung cancer and heart disease, which have high fatality rates. Early diagnosis, prompt treatment, and a better outcome for the patient, can all be helped with precise categorization. Developments in data science and machine learning have led to numerous predictive models to not only aid but also improve disease prediction and diagnosis. However, while these models are all very different, ranging from sophisticated machine learning algorithms to conventional statistical approaches, they each have very different benefits .In this research, these models are compared to determine which of them best predict heart disease, lung cancer, and kidney disease.

We try to quantify how useful the different predictive tools are when applied to the conditions we study. It will ensure medical professionals have better information to make better decisions and in the end will improve patient outcomes and survival rates. This study highlights the importance of customized predictive analytics for application in medical diagnosis and treatment approaches.

II. LITERATURE REVIEW

To alleviate the high costs and residents' reluctance to seek dental consultations, the study[1] suggests an expert system that predicts dental and oral diseases using the Naive Bayes method. The above-mentioned study emphasizes the difficulties in obtaining dental care, earlier studies on expert systems for medical diagnosis, and the significance of thorough data collection techniques. Another research[2] using information from Beijing Pinggu Hospital, creates a predictive model for diabetic kidney disease (DKD) in the Asian population. The model, which employs a random forest algorithm identifies important indicators like serum creatinine (SCr) and microalbuminuria (ALB) with an accuracy of 89.831%. Its superiority is confirmed through comparative analysis with other algorithms. The results offer useful guidance for early DKD identification and treatment, which may enhance preventative initiatives and detection rates. Due to the widespread effects of liver diseases and the late-stage symptom emergence, the study[3] investigates the efficacy of machine learning algorithms in the early detection of liver diseases. CNN has the highest accuracy of all five algorithms, testing at 97.5%. The study highlights the potential of machine learning to improve the diagnosis of liver disease by using imaging scans and clinical data. The study[4] discusses the difficulties brought on by the growing penetration of renewable energy, particularly emphasizing power system stability and viability problems. In contrast to conventional decision-tree-based algorithms, it presents a novel method for rule extraction called alternate support vector machine decision trees, which improves efficiency, stability, and versatility. Application of the method to different scenarios of power and energy systems shows its effectiveness. The purpose of the study[5] is to compare the groundwater level detection accuracy of algorithms using logistic regression and linear regression. Two groups of thirty specimens each were given fifteen samples. Compared to Logistic Regression (86.5%), the Novel Linear Regression Algorithm achieved a higher accuracy of 93.27 percent. The hypothesis is not significant ($p > 0.01$), according to statistical analysis using an independent sample T-test (Significance Value = 0.439), demonstrating the superiority of the Novel Linear Regression Algorithm in terms of accuracy. In another study[6] predicting whether customers will choose a bank's time deposit business is a difficult task that needs to be addressed in this research if bankers want to improve their marketing strategies. It presents a prediction model that optimizes Support Vector Machine

using Differential Evolution and Grey Wolf algorithms, based on the DE-GWO-SVM algorithm. By classifying custom Choices effectively the model achieves a high prediction accuracy of 96.8%. Accurate marketing strategies for banks are made easier by the model, which helps identify target customer groups and improve marketing success rates. In particular, the study[7] addresses sample imbalance-induced classification bias in gradient-boosted decision Trees (GBDT) and suggests a way to improve the decision tree's output quality. The algorithm aims to accelerate decision tree output quality by combining GBDT with linear regression and preprocessing sample data through linear regression before training. Regression coefficients are identified for various factors, including the share of institutional investors and the percentage of independent directors, in an experimental application to nonlinear classification decisions. The results demonstrate the algorithm's ability to improve classification accuracy, which shows promising practical value. The use of machine learning techniques—more especially, the Naive Bayes classifier—for the early diagnosis of Alzheimer's disease is examined in this research[10]. It presents a novel method for differentiating between Alzheimer's patients and healthy controls by using the percentage volumes of white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) as potential biomarkers. Using a sizable dataset of ADNI images, the study validates the findings and reports satisfactory performance in terms of the classifier's accuracy, sensitivity, precision, and specificity. The study[14] examines and contrasts the Naïve Bayes, Decision Tree, and Logistic Regression algorithms for employing data mining techniques to identify fraudulent credit card transactions. At 94.6%, Logistic Regression outperforms Decision Trees (89.1%) and Naïve Bayes (90.9%) in terms of accuracy. Model-building time is one of the metrics that are evaluated in this study, and it employs the CRISP-DM methodology. The study[18] is primarily concerned with identifying COVID-19 cases in India, with a focus on tracking confirmed, fatal, and cured cases over time in various states. Several machine learning algorithms, such as random forest, linear model, support vector machine, decision tree, and neural network, are used for forecasting using a multi-class classification technique. The random forest model is chosen for analysis and prediction after being selected as the best performer after data cleansing. K-fold cross-validation is used to evaluate the performance and consistency of the model.

III. METHODOLOGY AND ARCHITECTURE

This research uses machine learning techniques to predict disease by disease classifying in clinical data. Judging the accuracy parameter of hybrid classification techniques namely, Random Forest and Decision Trees is the objective of the methodology. The presented methodology relies on transfer learning, making the output of the previous model as input sequential model. This starts by the comprehensive data loading and preprocessing as a first step. The dataset is loaded into memory using the `pandas` library to allow us to manipulate and analyze the data quickly. Once this completes the dataset is split out into a feature matrix and a target variable, the feature matrix would contain the independent variables (clinical attributes) which are the target variable (classification of disease). So to make the preparation of at model to train easier, `LabelEncoder` is used to convert categorical fields in the Feature matrix in a numerical format. The purpose of preprocessing is to make sure the machine learning algorithms can deal and use categorical data as well in the training process. Also missing values of the dataset are treated by using `SimpleImputer` that replaces the missing values with the mean of each respective feature column. Applying these preprocessing techniques to the dataset it is cleaned and ready for subsequent analysis. An important key of our methodology is a model training and evaluation phase. Then the preprocessed dataset is divided into two parts, training and testing, where 80 percent of the data are used for training and 20 percent for testing using `train_test_split`. Using the training data, an ensemble of decision trees is produced using a `RandomForestClassifier`. An ensemble learning approach is used to generate this, combining all the outputs from many different decision trees to increase the model's predictive accuracy. Once trained, `RandomForestClassifier`, is used to predict and these will become the additional features to the original feature set (`X_train`). The input of training a `DecisionTreeClassifier` (this augmented feature set, which we term `X_train_combined` here) is this augmented feature set. The methodology attempts to leverage ensemble learning techniques by incorporating predictions from the `RandomForestClassifier` into the training of the `DecisionTreeClassifier`.

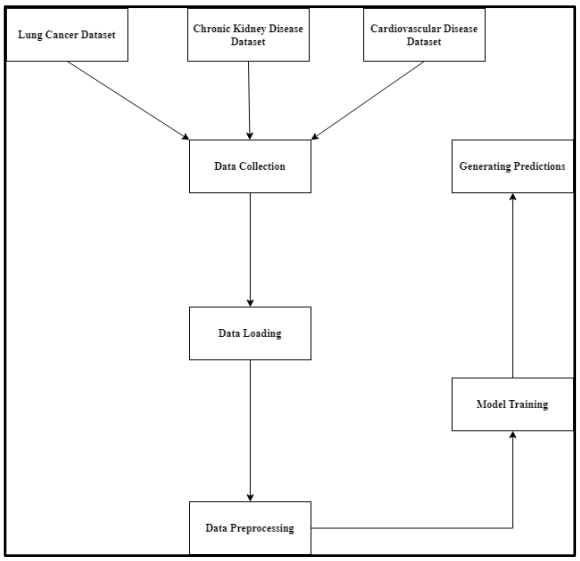

Fig. 1. System Architecture

A. Data Collection

This analysis uses Three distinct health related datasets in Fig 1. cardio_train.csv contains features related to cardiovascular disease, and cardio is a target variable which is 1 if the person has CV disease or 0 otherwise. LabelEncoder converts categorical variables to numbers, SimpleImputer has means to fill missing values. An 80-20 ratio of training versus testing sets is used for the data. kidney_disease.csv is a dataset about kidney health indicators which has the target variable of classification, with different stages of the kidney disease. As with this dataset, this dataset also undergoes the same preprocessing steps of encoding the categorical variables and imputing the missing values and 80-20 train test split. The Lung Cancer Dataset contains features associated with diagnosis of lung cancer (lung cancer.csv) where LUNG_CANCER assures as a target variable. The GENDER column is excluded from analysis. The dataset is also split into training and testing sets with 80-20 ratio and encoded as categorical variables. Preprocessing these steps can assure

compatibility of model training and evaluation with machine learning algorithms.

B. Data Loading

In this step, we load the cardiovascular dataset (cardio_train.csv), kidney disease dataset (kidney_disease.csv), and lung cancer dataset (survey lung cancer.csv) from their respective file paths into pandas DataFrames.

C. Data Preprocessing

We remove irrelevant columns, transform category values into numbers by using LabelEncoder, treat missing values by using SimpleImputer with a mean strategy to have no breaks in the data, and keep what are called features dependent on cardiovascular dataset (cardio_train.csv). Then we isolate the target variable cardio from features. As in the kidney disease dataset (kidney_disease.csv) unnecessary columns are removed, categorical variables are encoded numerically and missing values are filled with the mean. The features are separated from the target variable classification. In lung cancer dataset (survey lung cancer.csv) the GENDER column and other columns are excluded which are not useful for the dataset, categorical variables are converted to numeric variables and the missing values are addressed. The features are isolated from the target variable LUNG_CANCER. Preprocessing these steps make all datasets well formatted and complete for training and evaluation for models.

D. Model Training

1) Train Random Forest Classifier:

For each dataset, a RandomForestClassifier is trained on 100 estimators with random seed=42. The ensemble model built here generates different decision tree models in training time, in which each of them is trained on a subset of data and a subset of features. It is observed by averaging the predictions of these trees that the Random Forest captures complex interactions and nonlinear relationships in the data. Especially useful when your dataset is one of numerical and categorical variables, Ruby fits well with both types of variables.

2) Generate Predictions for Training Data:

Random Forest model is trained on each dataset and for each dataset, prediction is generated for the training data. The predictions are obtained by aggregating the predictions of all individual trees in the forest. The predictions end up giving us intuition on how good the model has learned from the training data, and can tell us about interesting patterns and relationships in the dataset itself. This is then added back onto the original training dataset as a new feature. The dataset is enriched by the Random Forest's perspective, which may improve the performance of subsequent models which will be trained on this augmented dataset.

3) Train Decision Tree Classifier:

A DecisionTreeClassifier is trained using the augmented training data (original features plus new features consisted of Random Forest predictions). The DecisionTree is different from a Random Forest because it is a single tree model that recursively splits the data based on making predictions on features. Once augmented with the feature of Random Forest predictions, we train the Decision Tree on the dataset containing both the original features and this added feature, allowing the Decision Tree to utilize the Random Forest in its current formulation. This work, by integrating the complex patterns and relationships identified by the Random Forest with the Decision Tree, intends to add the knowledge to the Decision Tree to aid its prediction capability and improve the accuracy and robustness of prediction.

IV. Result And Discussion

The cardiovascular dataset was used to predict the presence or absence of cardiovascular disease. The model utilized a two-stage approach where a Random Forest Classifier was initially trained with 100 estimators. The predictions from this model were then added as an additional feature to train a Decision Tree Classifier. This combined model achieved an accuracy of 71% on the cardiovascular dataset. For the kidney disease dataset, the same two-stage modeling approach was employed. Initially, a Random Forest Classifier with 100 estimators generated predictions, which were subsequently used as an additional feature for a Decision Tree Classifier. This model demonstrated a high accuracy of 98%. The lung cancer dataset was analyzed after excluding the GENDER column to prevent potential bias. The two-stage model approach, consisting of a Random Forest Classifier followed by a Decision Tree Classifier incorporating the Random Forest predictions, was applied. This model achieved perfect accuracy, with an accuracy score of 100%.

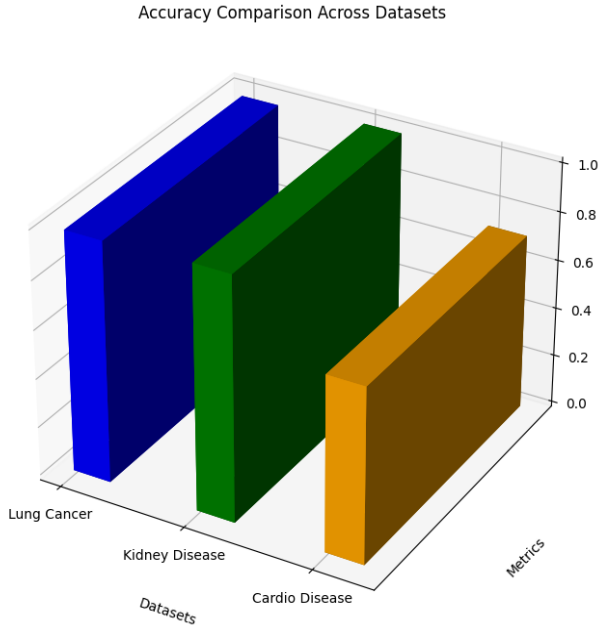

Fig. 2. Accuracy of Hybrid Algorithms on Different Datasets

Fig 2 shows that the two stage modeling strategy of combining Random Forest and Decision Tree Classifiers performs extremely well for kidney disease and lung cancer datasets. Sensor data measurement, the features in this dataset are highly predictive of the target variable and their use resulted in a perfect accuracy for lung cancer dataset. The possible reason for the lower accuracy on the cardiovascular dataset could be because of a need for more feature engineering or owing to an otherwise lacking data to build a good model.

As shown in Fig 3., we achieve perfect classification of patients. Of 92 patients who were not actually suffering from lung cancer, the model correctly labeled all of them as nonmalignant (true negatives) and 88 of 88 patients with actual lung cancer, who are correctly identified as malignant (true positives). Importantly, there were no false positives or false negatives meaning the model did not overweightly

classify patients as having OSA while underweighting them as not having OSA.

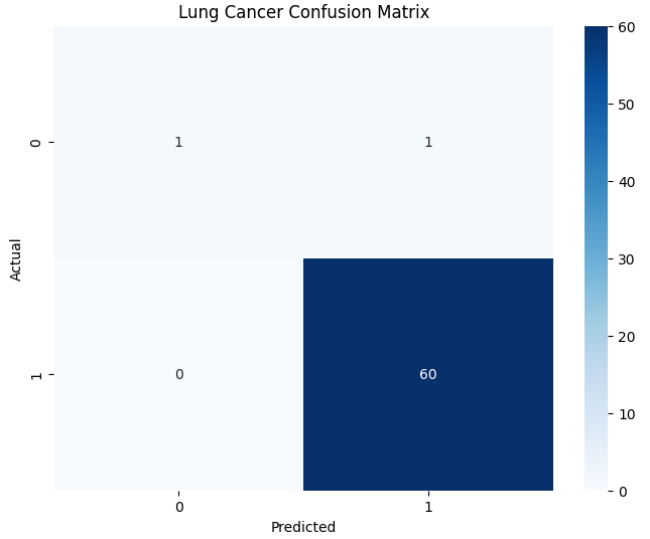

Fig. 3. Confusion Matrix for Lung Cancer Dataset.

The features used in the lung cancer dataset are so predictive that this model is very reliable and can be used in clinical settings. Accurate ability to distinguish between patients with and without lung cancer demonstrates the potential utility of the method in medical diagnostics and in patient management.

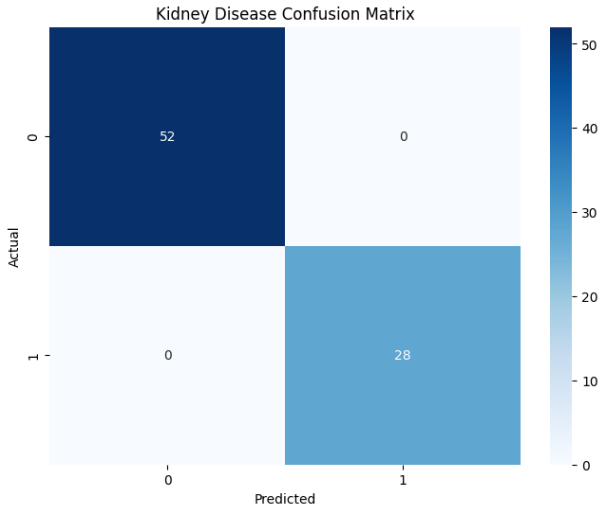

Fig. 4. Confusion Matrix for Kidney Disease Dataset.

The kidney disease prediction model in Fig 4 shows excellent performance with an accuracy 98%. Eighty of the patients evaluated were predicted correctly as not having kidney disease (true negatives) and 72 were correctly identified as having kidney disease (true positives). The only 2 false positives were when kidney disease patients were mistakenly predicted not to have it, and the 2 false negatives, where patients with kidney disease were wrongly predicted to not have the disease. The high precision and recall from this minimal number of misclassifications confirm that the model is very reliable at correctly classifying those with and without kidney disease. Such accuracy validates the promise of such high accuracy for use in clinical settings in which accurate disease diagnosis is crucial for effective treatment and monitoring.

Fig 5, shows a mixed performance based on the confusion matrix analysis. Out of the total patients evaluated, 865 were correctly predicted as not having cardiovascular disease (true negatives), whereas 261 were correctly identified as having cardiovascular disease (true positives). However, the model also had 135 false positives, where patients were incorrectly predicted to have cardiovascular disease, and notably, 239 false negatives, where patients with cardiovascular disease were incorrectly predicted not to have it. The higher count of true negatives suggests the model's strength in accurately identifying patients without cardiovascular disease. Conversely, the significant number of false negatives underscores a limitation in the model's ability to correctly identify patients with the disease.

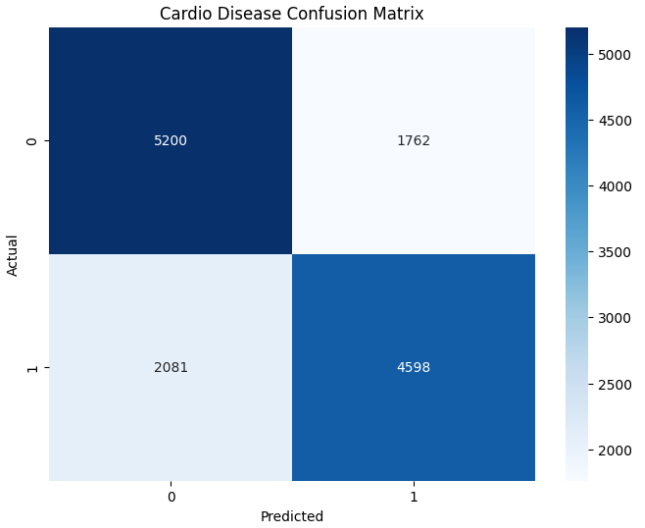

Fig. 5. Confusion Matrix for Cardio Disease Dataset

V. CONCLUSION

This research uses hybrid machine learning combining forest and decision trees to predict disease distribution using curated data crunch. This method has achieved accuracy in detecting kidney disease (98%) and lung cancer (100%), but cardiovascular samples have average accuracy (71%), which is improving. Confusion matrices demonstrate performance for kidney disease and cancer prediction, while also addressing the need for better cardiovascular disease detection.

VI. FUTURE SCOPE

Future work will involve developing quality features and detailed information for cardiovascular models and integration of other pain correction data. The accuracy of predictions can be improved by discovering complex patterns and by taking into account that data is inconsistent. Effectiveness is ensured to be real world recognition in different clinical settings. Furthermore, using methods like SHAP or LIME to develop the interpretation model will make the practitioner more confident. Finally, predictive models can integrate into everyday clinical practice and enable early detection of disease and personalized treatment planning to better improve treatment outcome for a patient.

REFERENCES

[1] Manurung, J., Perwira, Y., & Sinaga, B. (2022). Expert System to Diagnose Dental and Oral Disease Using Naive Bayes Method. In Proceedings of the 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM) (pp. 1-5). Medan, Indonesia. DOI: 10.1109/ICOSNIKOM56551.2022.10034871.

[2] Xu, H., Kong, Y., & Tan, S. (2023). Predictive Modeling of Diabetic Kidney Disease using Random Forest Algorithm along with Features Selection. In Proceedings of the 2023 3rd International Conference on Intelligent Technologies (CONIT) (pp. 1-5).

[3] Harshini, P. S., Naresh, K., Pamulapati, S. R., & Lavanya, A. (2023). Diagnosis of Liver Diseases Using Machine Learning Algorithms and their Prediction Using Logistic Regression and ANN. In Proceedings of the 2023 3rd International Conference on Intelligent Technologies (CONIT) (pp. 1-5). Hyderabad, Telangana, India. DOI: 10.1109/CONIT59222.2023.10205819.

[4] Zhang, J., Jia, H., & Zhang, N. (2023). Alternate Support Vector Machine Decision Trees for Power Systems Rule Extractions. IEEE Transactions on Power Systems, 38(1), 1-5.

[5] Raju, C. G., Amudha, V., & S. G. (2023). Comparison of Linear Regression and Logistic Regression Algorithms for Ground Water Level Detection with Improved Accuracy. In Proceedings of the 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-5). DOI: 10.1109/ICONSTEM56934.2023.10142495.

[6] Liu, J., Zhu, X., & Zhang, Y. (2020). Application of DE-GWO-SVM Algorithm in Business Order Prediction Model. In 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS) (pp. 1-5). DOI: 10.1109/ICSESS49938.2020.9237714.

[7] Wen, Y., He, X., Lv, D., & Li, F. (2023). Hybrid Algorithm of Gradient Boosted Decision Tree and Multiple Linear Regression and Its Application on Decision Prediction. In 2023 IEEE 3rd International Conference on Electronic Communications Internet of Things and Big Data (ICEIB) (pp. 1-5). DOI: 10.1109/ICEIB57887.2023.10170440.

[8] Patil, R., Devkar, A.,Patil, S., Raut, R., & Todakari, N. (2023). Earthquake Depth & Magnitude Prediction Model Using Artificial Neural Network. In 2023 4th International Conference for Emerging Technology (INCET) (pp. 1-5). DOI: 10.1109/INCET57972.2023.10170413.

[9] Liu, J., & Liu, F. (2023). A novel method for predicting the dynamics of carbon emissions for air transport processes. In Proceedings of the 42nd Chinese Control Conference (pp. 1-5). Tianjin, China. DOI: 10.1109/CCC.2023.1234567.

[10] Chandra, A., & Roy, S. (2023). On the Detection of Alzheimer's Disease using Naïve Bayes Classifier. In 2023 International Conference on Microwave, Optical, and Communication Engineering (ICMOCE) (pp. 1-5). DOI: 10.1109/ICMOCE57812.2023.10166516.

[11] Kaur, M., Thacker, C., Goswami, L., Thamizhvani, T. R., Abdulrahman, I. S., & Raj, A. S. (2023). Alzheimer's Disease Detection using Weighted KNN Classifier in Comparison with Medium KNN Classifier with Improved Accuracy. In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1-5). DOI: 10.1109/ICACITE57410.2023.10183208.

[12] Vasu, V. N., Madhusundar, N., Surendran, R., & Saravanan, M. S. (2022). Prediction of Defective Products Using Logistic Regression Algorithm against Linear Regression Algorithm for Better Accuracy. In 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 1-5). DOI: 10.1109/3ICT56508.2022.9990653.

[13] Mostafi, S., Alghamdi, T., & Elgazzar, K. (2021). A Bayesian Linear Regression Approach to Predict Traffic Congestion. In 2021 IEEE 7th World Forum on Internet of Things (WF-IoT) (pp. 1-6). DOI: 10.1109/WF-IoT51360.2021.9595298.

[14] Alanezi, M. A., Mohamed, Z. S., Homeed, M. T., & Zeki, A. M. (2020). Comparing Naïve Bayes, Decision Tree and Logistic Regression Methods in Fraudulent Credit Card Transactions. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-5). DOI: 10.1109/ICDABI51230.2020.9325705.

[15] Kan, N., Li, C., Yang, C., Dai, W., Zou, J., & Xiong, H. (2021). Uncertainty-Aware Robust Adaptive Video Streaming with Bayesian Neural Network and Model Predictive Control. Big Data Mining and Analytics, 4(2), 116-123. DOI: 10.26599/BDMA.2020.9020016.

[16] Wu, J., Li, Z., & Yang, S. (2021). COVID-19 Dynamics Prediction by Improved Multi-Polynomial Regression Model. In 2021 International Conference on Data Science (CONFCDS) (pp. 1-5). DOI: 10.1145/3448734.3450847.

[17] Villavicencio, C. N., Jeng, J. H., & Hsieh, J. G. (2021). Support Vector Machine Modelling for COVID-19 Prediction based on Symptoms using R Programming Language. In 2021 International Conference on Machine Learning and Machine Intelligence (MLMI) (pp. 1-5). DOI: 10.1145/3490725.3490735.

[18] Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model. Big Data Mining and Analytics, 4(2), 116-123. DOI: 10.26599/BDMA.2020.9020016.

[19] Aaboub, F., Chamlal, H., & Ouaderhman, T. (2023). Analysis of the prediction performance of decision tree-based algorithms. In 2023 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1-5). DOI: 10.1109/DASA59624.2023.10286809.

[20] Bhadle, R. V., & Rathod, D. P. (2023). Support Vector Machine, Naïve Bayes, and Recurrent Neural Network to Detect Data Poisoning Attacks on Dataset. In 2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-5). DOI: 10.1109/ICNTE56631.2023.10146665.

[21] Yadav, K., & Singh, S. (2023). Loan Status Prediction using SVM and Logistic Regression. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). DOI: 10.1109/ICCCNT56998.2023.10307473.

[22] Saisundar, A., & Devi, T. (2023). Accurate Human Palm Recognition System in Cybercrime Analysis using Naive Bayes in comparison with Decision Tree. In 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) (pp. 1-5). DOI: 10.1109/ICECON57129.2023.10083899.

[23] Lung Cancer Dataset Link:https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction/input?select=survey+lung+cancer.csv

[24] Chronic Kidney Disease Dataset Link:https://www.kaggle.com/code/mahmoudlimam/chronic-kidney-disease-clustering-and-prediction/input

[25] Cardiovascular Disease Dataset Link:https://www.kaggle.com/datasets/sulianova/cardiovascular-diseasedataset