iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

DETECTING CHRONIC DISEASES: A LOGISTIC REGRESSION APPROACH

Anjali Pralhad Landge
M.Tech Student, Artificial Intelligence and Data Science, CSMSS CSCOE Chh.Sambhajinagar India

Dr Ashwini S.Gavali
Professor, Artificial Intelligence and Data Science, CSMSS, CSCOE ,Chh.Sambhajinagar,India

Prof. Khan Mahmood Ahmed
Assistant Professor, Artificial Intelligence and Data Science,
CSMSS CSCOE,Chh.Sambhajinagar,India

ABSTRACT
This Diabetes Mellitus and Chronic Kidney Disease (CKD) are among the most critical chronic health conditions worldwide, requiring early and accurate diagnosis to reduce complications and mortality. Traditional diagnosis often depends on clinical expertise and laboratory interpretation, which may delay early-stage detection. This experimental study presents a machine learning-based predictive framework for the identification of diabetes and kidney-related ailments using structured and semi-structured healthcare data. The proposed methodology integrates data collection, preprocessing, feature selection, normalization, classification, and performance evaluation to improve disease prediction accuracy. In the preprocessing phase, missing values, irrelevant symbols, and inconsistencies are handled to enhance data quality. A feature selection mechanism based on relevance and redundancy is incorporated to identify the most informative attributes and reduce computational complexity. The proposed prediction model employs Sparse Multinomial Logistic Regression along with machine learning-based classification strategies for effective disease categorization. Experimental validation is carried out using benchmark datasets including the PIMA Indians Diabetes Dataset and Chronic Kidney Disease-related data. Performance is evaluated using precision, recall, accuracy, F-score, average correlation, and Area Under the Curve (AUC). The obtained results demonstrate that the proposed approach achieves superior predictive performance, with accuracy values of 96.08% for diabetes prediction and 96.25% for kidney disease prediction, outperforming conventional baseline techniques. The reduced average correlation also indicates lower redundancy among selected features, thereby improving model robustness and interpretability. The findings confirm that machine learning-assisted clinical prediction can serve as a reliable decision-support tool for the early diagnosis of chronic diseases. This work contributes toward the development of intelligent, data-driven healthcare systems capable of supporting physicians in timely and effective disease management

Keywords:
Diabetes, Kidney, Kernel Techniques

INTRODUCTION

Diabetes and its Types

There are numerous underlying causes for the metabolic disorder known as diabetes mellitus (DM). The condition is characterised by persistent hyperglycemia and modifications in how proteins, lipids, and carbohydrates are metabolised as a result of insulin deficiency, insulin action, or both. Diabetes is one of the chronic diseases. Diabetes can damage the neurons and blood arteries in the kidneys, heart, lower legs, and eyes if it is not adequately controlled. Problems could develop if blood glucose levels are high for an extended period of time. Gum disease and tooth decay are two examples of oral issues. Diabetic retinopathy can, in severe cases, cause blindness and vision loss. Cardiovascular diseases (CVD) (insufficient blood flow to the feet and legs) are a group of heart and blood vessel illnesses that include peripheral artery disease, heart attacks, and strokes. Kidney function may be compromised or absent in patients with kidney disease, often known as diabetic nephropathy. [7] The three different kinds of diabetes include type 1 diabetes, type 2 diabetes, and gestational diabetes.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Type 1 diabetes (T1D)
Those who have type 1 diabetes do not produce enough insulin in their bodies. Without insulin, bodily cells must use alternate energy sources as they are unable to absorb glucose from the bloodstream. An excess of glucose in the blood causes diabetes and its effects. This type of diabetes (IDDM) is also known as insulin-dependent diabetes mellitus. It can affect anyone at any age, though teens and young adults are the ones it commonly affects. A careful balancing act is required between insulin injections (and, in some circumstances, oral drugs), exercise, meal planning, and lifestyle changes. Frequent urination, unusual thirst, unusual hunger, rapid weight loss, weakness and exhaustion, nausea, and irritability are all signs of type 1 diabetes.
Type 2 diabetes (T2D)
In people with type 2 diabetes, insulin resistance can range from primarily present to predominately absent. Although the pancreas produces insulin, it might not be enough to keep blood glucose levels within normal ranges or the cells might be insulin-resistant. Teenagers and young toddlers are now becoming more prone to the illness, despite the fact that people over 40 are more likely to contract it. Type 1 diabetes symptoms include drowsiness, dry, itchy skin, unwelcome weight gain or loss, blurred vision, tingling, numbness, soreness in the lower legs, ease of weariness, poor wound or scratch healing, and recurrent infections (such as vaginal infections). Combinations of factors, such as food, exercise, lifestyle changes, and occasionally oral medicines or insulin, are needed to treat type 2 diabetes. [6].
Gestational diabetes
Elevated blood glucose (sugar) levels in pregnant women without a history of diabetes are referred to as gestational diabetes. It affects 2% to 4% of all expectant mothers and usually goes away after delivery. Women with a history of gestational diabetes are more likely to develop type 2 diabetes. There is no known aetiology for this type of diabetes. The placenta helps the baby grow, but hormones from the placenta also prevent the mother's body from properly using her own insulin, which causes insulin resistance. When a mother's body is unable to create and use all of the insulin required during pregnancy, gestational diabetes develops. The majority of women are unaware of any signs or symptoms of gestational diabetes. An increase in hunger and frequent urination are two symptoms.
Diabetes mellitus can be lethal if untreated, however early detection can significantly lower the risk. Numerous medical diagnostic methods are already in use for early diagnosis. Early risk projections may be provided by machine learning algorithms. In a recent study, forecasting the risk of diabetes mellitus had good results. In the discipline of machine learning, algorithms are used to teach computers without the assistance of people. Without having to directly programming them, we may train them to complete a task and then use that training to tackle related tasks. Accuracy is a persistent problem in medical science, as different algorithms may offer varying degrees of accuracy on the same data set.
To name a few machine learning and classification methods, naive Bayes classifiers, support vector machines, decision trees, random forests, and artificial neural networks (ANNs) are good at predicting risk. Because of the algorithms' capabilities for computing and data management, this is possible. Measures of classification accuracy can be used to select the optimum algorithm and classification accuracy. However, this amount is insufficient to precisely and successfully select the best course of action. When selecting the best conclusion, additional variables such as the receiver operating characteristic (ROC) value, F-score, and computation time should be taken into account. Calculation time, ROC value, F-score, and classification accuracy are among the metrics. Future researchers will benefit from the findings of this study in creating a standard procedure for categorising DMs.

About Chronic Kidney Diseases
Machine Learning Algorithms
Machine learning (ML), a rapidly increasing field, is being used in a variety of medical applications. All ML models make predictions based on historical data using that data. Diabetes detection will become much simpler and less expensive due to recent advances in ML. There are numerous diabetic data sets accessible. Therefore, ML is required for medical diagnosis. The purpose of this study is to forecast a patient's risk of developing diabetes. Algorithms for machine learning are employed. There are two distinct learning modes for the study.
1) Supervised
2) Un-Supervised.
The goal of a supervised learning algorithm is to predict using labelled data. In supervised learning, the information is labelled. It serves as an example of what a student might learn in a simulation from an instructor. Unsupervised learning, on the other hand, does not label the data. Given that it is founded on existing

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

information, it more closely resembles self-learning. Predicting a variable's value is the goal. The information is displayed as a collection of traits and characteristics. Directed learning already has a known outcome. Decision trees (DT), random forests, linear regression, logistic regression, naive Bayes classifiers, k-nearest neighbours (k-NN), support vector machines (SVM), and artificial neural networks (ANN) are a few of the techniques that are frequently utilised.

In unsupervised learning, the outcomes are not predetermined, and the data is made up of values only. On the basis of self-learning, the model makes predictions. These models' primary goals are to forecast, categorise, detect, segment, and categorise data. Information retrieval, image analysis, biology, data compression, and computer graphics are a few examples of applications for machine learning.

RELATED WORK

Birjais et al.'s research[8] made use of the PIMA Indian Diabetes (PID) data collection. It has 768 instances and 8 characteristics, and it can be found in the UCI machine learning repository. The World Health Organisation (WHO) listed diabetes as one of the chronic illnesses with the highest rate of rise in 2014. They want to focus more on diabetes diagnosis. Gradient boosting, logistic regression, and naive Bayes classifiers were used to determine whether a person had diabetes or not. The accuracy of logistic regression was 79%, that of naive Bayes was 77%, and that of gradient boosting was 86%.

Sadhu, A., and Jadli, A. conducted an experiment using diabetes data from the UCI repository [9]. In total, there were 520 instances of each of the 16 traits. They tried to concentrate their efforts on detecting diabetes early. To validate the employed data set, seven classification techniques were applied: k-NN, logistic regression, SVM, naive Bayes, decision trees, random forests, plus multilayer perceptrons. With an accuracy score of 98%, the random forests classifier emerged as the most accurate machine learning model for the pertinent data set, which was followed by logistic regression at 93%, SVM at 94%, naive Bayes at 91%, decision tree at 94%, random forests at 98%, as well as multilayer perceptron at 98%.

The diabetes data set utilised in the study by Xue et al. [10] was gathered from the UCI repository and contained 520 patients and 17 variables. They concentrated on the early detection of diabetes. They trained on the real data of 520 diabetic patients and those who were at risk of developing diabetes, aged 16 to 90, using supervised machine learning techniques like SVM, naive Bayes classifiers, and LightGBM. The SVM performs best when comparing classification and recognition accuracy. The most used classification algorithm is the naive Bayes classifier, which has an accuracy rating of 93.27%. SVM has the highest accuracy rate at 96.54%. LightGBM's accuracy is just 88.46%. This demonstrates that the best classification technique for predicting diabetes is SVM.

Le et al.'s[11] tested with forecasting the propensity for early-stage diabetes to develop. The 520 patients and 16 variables that made up the data set used in this investigation were taken from the UCI repository. They suggested utilising machine learning to forecast patients' early onset of diabetes. In order to reduce the number of input characteristics needed, the multilayer perceptron (MLP) was improved utilising a unique wrapper-based feature selection method that also made use of the grey wolf optimizer (GWO) and adaptive particle swarm optimisation (APSO). Additionally, they compared the results of this strategy with those of a number of traditional machine learning techniques, including SVM, DT, k-NN, NBC, RFC, and logistic regression (LR).

With LR, a 95% accuracy rate was reached. SVM had a 95% accuracy rate, NBC had a 93% accuracy rate, DT had a 95% accuracy rate, and RFC had a 96% accuracy rate. In addition to requiring fewer features, the computational results of the proposed methodologies show that higher prediction accuracy may be achieved (97% for APSO-MLP and 96% for GWO-MLP). This research could be helpful to physicians, other healthcare workers, and clinical practises.

Julius et al.[12] examined a data set obtained from the UCI repository using the Waikato Environment for Knowledge Analysis (Weka) application platform. There were 520 samples in the data set, and each sample had a set of 17 attributes.

The goal of this work was to use machine learning classification approaches based on observable sample attributes to predict diabetes at an early stage. The k-NN, SVM, functional tree (FT), and RFCs were utilised as classifiers. K-NN (98%) had the highest accuracy, followed by SVM (94%), FT (93%), and RF (97%).

Findings from Shafi et al.[13] show that diabetes, a serious illness, is always challenging to diagnose early. In this study, machine learning-based categorization approaches were utilised to build a model that could handle any problem and be applied to early diabetes detection. The authors of this study put a lot of effort into establishing a model that could accurately predict a patient's likelihood of acquiring diabetes. In this work, three ML approach classification algorithms—DT, SVM, and NBC—were looked at and assessed according to a variety of standards. The PID data set from the UCI library was used in the study to speed up the procedure and

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

get precise results. The experimental results demonstrated that the accuracy of the NBC approach, which was 74%, was sufficient, followed by the accuracy of the SVM, which was 63%, and the accuracy of the DT, which was 72%. The created framework and the used ML classifiers may one day be used to identify or diagnose new diseases. The study might be enhanced and extended upon for the study of diabetes, and the researchers sought out alternative approaches with gaps in the data.

The study on the prognosis of diabetes illness by Khanam et al. Because diabetes has no known cure, early detection is essential. In this study, data mining, machine learning (ML), and neural network (NN) techniques were used to predict the presence of diabetes. They developed a technique for accurately predicting diabetes. They made use of the UCI repository's PID data gathering. Details about 768 patients and their nine characteristics were included in the data collection. The data set was subjected to seven machine learning (ML) approaches in order to predict diabetes: DT, k-NN, RFC, NBC, AB, LR, and SVM. They used the Weka programme to preprocess the data. They discovered a model that effectively predicts diabetes by combining LR and SVM. They created a NN model that included two hidden layers, many epochs, and they found that it had an accuracy of 88.6%. With a total score of 88.57%, ANN was ranked first, ahead of LR (78.15%), NBC (78.18%), and RFC (77.34%).

Sisodia et al. [15] used the PID data set that is kept in the UCI repository. In this data collection, there were 768 patients and 8 characteristics. They employed DT, SVM, and NBC, three ML classifiers, to identify people with diabetes. In comparison to the other models, NBC's accuracy was the highest (76.30%).

Agarwal et al.[16] also used the PID data set, which included 738 patient records, in their analysis. The authors tested the accuracy of this data set in identifying people with diabetes using a variety of models, including SVM, k-NN, NBC, ID3, C4.5, and CART. The SVM and LDA algorithms were the most accurate with an accuracy of 88%.

Rathore et al.[17] employed classification techniques like SVM and DTs to predict diabetes mellitus. The PID data collection provided the information for this investigation. PIMA India places a high premium on women's health. The SVM achieves 82% accuracy.

To predict diabetes mellitus, Hassan et al.'s[18] team used classification approaches such the DT, k-NN, and SVM. The SVM performed better than the DT and KNN algorithms, with a maximum accuracy of 90.23%.

Kandhasamy and Balamurali[19] investigated the J48, k-NN, RFC, and SVM prediction accuracy for the diabetic data set. Before preprocessing the data, the author discovered that the J48 technique had a higher accuracy than the others, at 73.82%. Preprocessing improved the accuracy of k-NN and RFC.

Meng et al.[20] investigated the J48, LR, and k-NN algorithms using the diabetes data set. The categorization accuracy for J48 was found to be the highest, at 78.27%.

Kumari and Chitra[23] used SVM, RFC, DT, MLP, and LR in addition to four k-fold cross-validations (k = 2, 4, 5, and 10). According to the researchers, MLP with four-fold cross-validation yields the best accuracy, with a score of 78.7%. They discovered that MLP outperformed every other algorithm.

The NBC, RFC, k-NN, SVM, DT, and LR algorithms were employed by Kavakiotis et al. [24] to predict diabetes. The algorithms were used via ten-fold cross-validation method. SVM had an accuracy rating of 84%, which was the greatest of all the approaches according to the analysis.

[25] AdaBoost, LogicBoost, RobustBoost, Naive Bayes, and Bagging were five machine learning (ML) algorithms that were explored in this study for the analysis and prediction of diabetes patients. A group of diabetic PIMA Indians were used as test subjects for the proposed remedies. With classification accuracy for the bagging and AdaBoost techniques of 81.77% and 79.69%, respectively, it was discovered that the estimated results were quite accurate.

A online application employing disease classifiers and actual data was suggested by Nai-Arun and Moungmai[21]. Between 2012 and 2013, 30,122 people in the twenty-six primary care clinics at Sawanpracharak Regional Hospital submitted the data for this component. Thirteen categorization models were examined in order to identify a prediction model before the web application was created. With the exception of the RFC algorithm, these models included the DT, NN, LR, NBC, and RFC algorithms, which all used bagging and boosting techniques. The accuracy and ROC curves of each model were constructed and contrasted with those of other models to ascertain how trustworthy each one was. The outcomes demonstrated that, in terms of accuracy and ROC curve, RFC was superior.

There are several potential causes for this. The RFC technique involved the random selection of both significant variables and data and input items. As a result, the accuracy values rose. This approach was picked to illustrate diabetes risk prediction, and the application was made using it as well.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Perveen et al.[26] used a data set gathered from the Canadian Primary Healthcare Sentinel Surveillance Network (CPCSSN) dataset to carry out their research. The study used the J48 (C4.5) DT as a base learner and independent data analysis methodology J48 to categorise patients with type 2 diabetes based on diabetes risk markers. AdaBoost and bagging ensemble methodologies were also used in the study. The CPCSSN used three separate ordinal adult groups for this categorization. The outcomes demonstrated that, in terms of overall performance, the AdaBoost ensemble technique outperformed bagged and a single J48 DT.

Mujumdar and Vaidehi[27] created a diabetes prediction model that included a few extrinsic factors that contributed to the onset of diabetes in addition to the usual components like glucose, BMI, age, insulin, and so on. The classification accuracy was enhanced when comparing the new data set to the old data set. The data set was treated to a variety of ML algorithms, and several classification techniques were used, with LR yielding the highest accuracy (96%). The AdaBoost classifier was determined to be the most dependable with a 98.8% accuracy rate. They used two separate sets of data to compare the efficacy of ML methods. When compared to the prior data set, it was clear that the model improved prediction of diabetes accuracy and precision.

Mercaldo et al.'s[28] technique of classifying diabetic patients was provided based on a set of variables chosen in compliance with WHO recommendations. using the latest recent machine learning techniques to analyse real data. The model was trained using six different categorization methods, and the accuracy and recall scores for the Hoeffding Tree method were 0.770 and 0.775, respectively. Information from the PIMA Indian community in Phoenix, Arizona was used to evaluate the strategy.

METHODOLOGY

This Data collecting comes first in the process. Data from both organised and unstructured sources is gathered by our suggested method. Data sets are divided into cleaning and test data sets when preprocessing is applied to the collected data. Finally, in order to increase the precision of the prediction findings, the training data set is trained using machine learning methods such as Sparse Multinominal Logistic Regression and neural network across a number of epochs. After the intended objective has been reached after several epochs, the generated model is prepared for testing. At this stage, the model is put to the test using a new set of data that was not used for training in order to assess how well it performs. The suggested model is suitable for deployment if it achieves the requisite accuracy in test data.

Data Collection

Real-world data consists of both structured and unstructured information, such as demographics, a patient's place of residence, and the results of lab tests. Structured information contains the patient's fundamental health information. In order to protect the patient's privacy, the data set does not include any of their identifying information, including name, ID, and location.

Preprocessing

Most structured data is preprocessed to account for the possibility of missing values. So, it is crucial to add the missing data, eliminate it, or change it in order to improve the quality of the data collection. The commas, punctuation, and white spaces are also removed during the preprocessing stage. When the data has undergone preprocessing, feature extraction and illness prediction are applied to it.

Model Description

The data set includes both organized and semistructured, as was already mentioned. The structured data includes tabularized information about the patient's living environment, laboratory test results, and the disease that they are suffering from, as well as demographic information about the patient's age, gender, height, weight, and other characteristics that are related to the disease's cause. The patient's medical symptoms and text-based information about the doctor interview make up the unstructured data. The prediction task benefits from the addition of unstructured input by obtaining more accurate results. 80% of the data set is used for training, while 20% is used for testing.

Disease Prediction Using Sparse Multinominal Logistic Regression

In order to forecast chronic illness, the suggested system employs the Sparse Multinominal Logistic Regression technique. The data set is first transformed into vector form, then word embedding is used to adopt zero values for the data's fill. After that, the convolution layer receives it.

The convolution layer provides the input to the pooling layer, which then performs the max pooling process. The fully connected layer receives the max pooling output before providing the classification outcomes to the output layer. The hospital's computer systems and internet archives have a large number of datasets available for all ailments. These datasets include various features for a variety of applications; not all of the traits are useful for detecting a particular ailment. Before using a system for the categorization of data, data which was before is

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

a crucial step. The model may produce false results if it was built on a data with incorrect entries. Similar to how not every attribute in a dataset equally contributes to detection. Including unnecessary features lengthens the model's processing time and reduces model performance. The performance of the classifier is severely impacted if all characteristics are employed in the prediction analysis. It depends on the doctor's experience to make a diagnosis of a disease from various symptom input data in hospitals and medical labs. The primary purpose of feature selection approaches is to examine the role that each characteristic plays in the output prediction process. Prior to building a model, selection of features is a crucial strategy for reducing data complexity by removing pointless and unnecessary elements.

The data matrix has been preprocessed and discretized with respect to the mean of each gene's expression (column). The number of output features (genes) say n is provided from outside by the user. The data matrix with classes c = {1, 2, · · ·,C} are the inputs. At the beginning, the first objective (obj1) i.e., the relevance of each gene is calculated by mutual information as per Equation 6. From the relevance score, the highest scorer gene id is extracted and added.

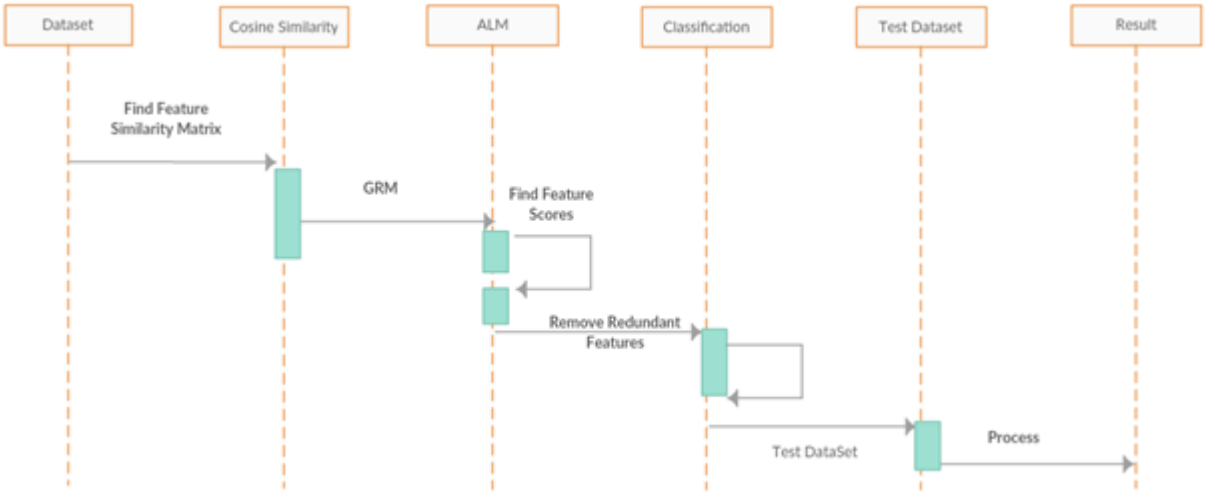


Figure 1.0 Sequence for proposed architecture

Algorithm 1 Proposed Feature Selection

Input: The feature id idle f t, first objective ob j1, second objective ob j2, |ob j1| = |ob j2| = |idle f t|.
Output: Non-dominated feature id idns, the second objective ob j2ns of non-dominated features.
1: k = 1;
2: for i = 1 : |idle f t| do
3: t = 0;
4: for j = 1 : |idle f t| do
5: if then(i! = j)
6: if then(ob j1(i) ≤ ob j1(j)&ob j2(i) ≤ ob j2(j));
7: else if then(ob j1(i) < ob j1(j)&ob j2(i) > ob j2(j)||ob j1(i) > ob j1(j)&ob j2(i) < ob j2(j));
8: else
9: t = 1;
10: break;
11: end if
12: end if
13: end for
14: if then(t == 0&j == |idle f t|)
15: idns(k) = i;
16: ob j2ns(k) = ob j2(i);
17: k = k + 1;
18: end if

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

19: end for
in the final solution set. Next a looping is performed for the remaining output features. Now the redundancy between the output feature and the remaining features (idle f t) is calculated as per Equation 5. If the output feature set contains more than one feature then the mean is considered as the redundancy score as in Equation.

$$\text{mean-redundancy}(i) = \sum_{k=1}^{F} (\text{mutual-info}[x_k, x_i])/|F|),$$

where F is output feature set, Xk is output feature vector and xi is the ith feature vector. Then the second objective (obj2) is modeled as the ratio of relevance to the redundancy and it is to be maximized. After calculating the two objectives for each feature the non-dominated features are identified. A reference feature is called the non-dominated feature if it satisfies the following conditions: 1) if the obj1 of the reference feature is greater than or equal to all the other futures' obj1 and the obj2 of the reference feature is greater than or equal to all the other features' obj2 2) if the obj1 of the reference feature is greater than all the other features' obj1 and the obj2 of the reference feature is less than all the other features' obj2 and vice-versa. Afterwards, from the non-dominated features, the feature having maximum obj2 is included in the output feature set.

RESULTS AND DISCUSSION

Service One real life data sets is used for the comparative study. The Diabetic Cancer dataset is collected from the website: www.biolab.si/supp/bi-cancer/projections/info/. The dataset contain two classes of samples.
1. Diabetic: Gene expression measurements for samples of Diabetic types and adjacent Diabetic tissue not containing type were used to build this classification model. It contains 50 normal tissue and 52 Diabetic type sample. The expression matrix consists of 12533 numbers of genes and 102 numbers of samples.
A unsupervised machine learning approach called tree is used to address classification issues. In this study, the primary goal of the decision tree is to forecast the target class using a decision rule derived from historical data. Nodes and internodes are used for categorization and prediction. Root nodes categorise the instances using various attributes. Although the leaf nodes indicate classification, the root nodes may contain two or more branches. The decision tree selects each node at each level by determining which attribute provides the most information gain overall [11].

Table 1.0 Attributes Summarization

Attribute	Abbreviation of Attributes
Number of times pregnant	pr
Plasma glucose concentration	pl
Diastolic blood pressure (mm Hg)	Pr
Skin fold thickness (mm)	Sk
2-Hour serum insulin (mu U/ml)	In
BMI (weightinkg/(heightinm) 2)	Ma
Diabetes pedigree function	Pe
Age in years	Ag
Class '0' or '1'	cl

Table 2.0 Key Index Parameters and Performance

Dataset	Methodology	Precision	Recall	Accuracy	FScore	Avg Corr	AUC
Diabetes UCI	Bayesian	96	90.23	93.22	93.20	0.32	95.92
PIMA Kaggle	Proposed Technique	98	94.23	96.08	96.80	0.23	98.92
Kidney Chronic Disease Dataset	Proposed Technique	97.435	98.70	96.25	98.0	0.21	99.21

Dataset for Pima Indians with Diabetes The Diabetes Dataset, known as PIDD [13], which is collected from the UCI Repository, is used to evaluate the suggested methodology. This dataset includes the medical information for 768 occurrences of female patients. The dataset also includes eight attributes with a numeric value, where a

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

value of class '0' is handled as a test result that was negative for diabetes while a value of class '1' is treated as a result that was positive for the disease.

The genuine data sets stated above are first standardised using the Min-Max normalisation approach. In order to discretize the data, each characteristic (gene) or column's average must first be determined. The number of output functions for all techniques in this article is taken to be 100. Using 10-fold cross-validation, the metrics of sensitivity, specificity, accuracy, and fscore score are calculated. The degree of overlap between the selected characteristics is then computed using the mean correlation. The correlation value decreases as the chosen functions become less redundant. The area under the ROC curve (AUC) is also displayed.

The metrics for the performance of the suggested method on several real datasets are shown in Table 1. The table shows that for the cancer data set, the diabetes method's sensitivity, specificity, AUC, and other metrics vary between 0.98, 0.9423, 0.9608, 0.9608, and 0.9892, all of which are better than the already employed approaches. Additionally, the suggested method's average correlation is 0.23, which is a bit lower than that of the other two approaches and demonstrates that the anticipated characteristics are least connected with the suggested strategy.

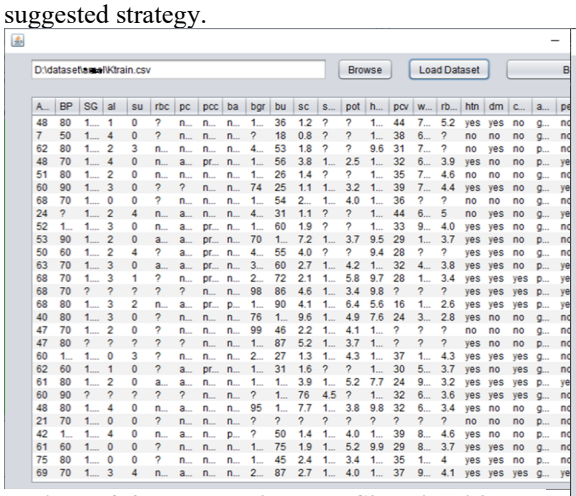


Figure 2.0 Data Parsing For Chronic Kidney Diseases

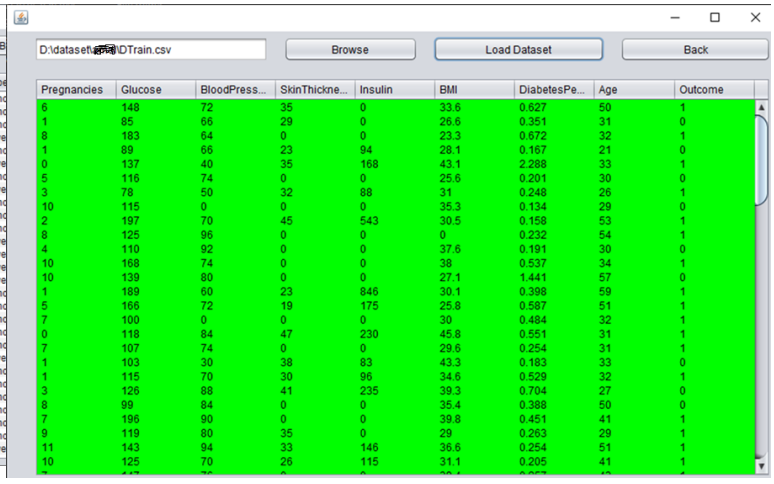


Figure 3.0 Data Parsing For Diabetes

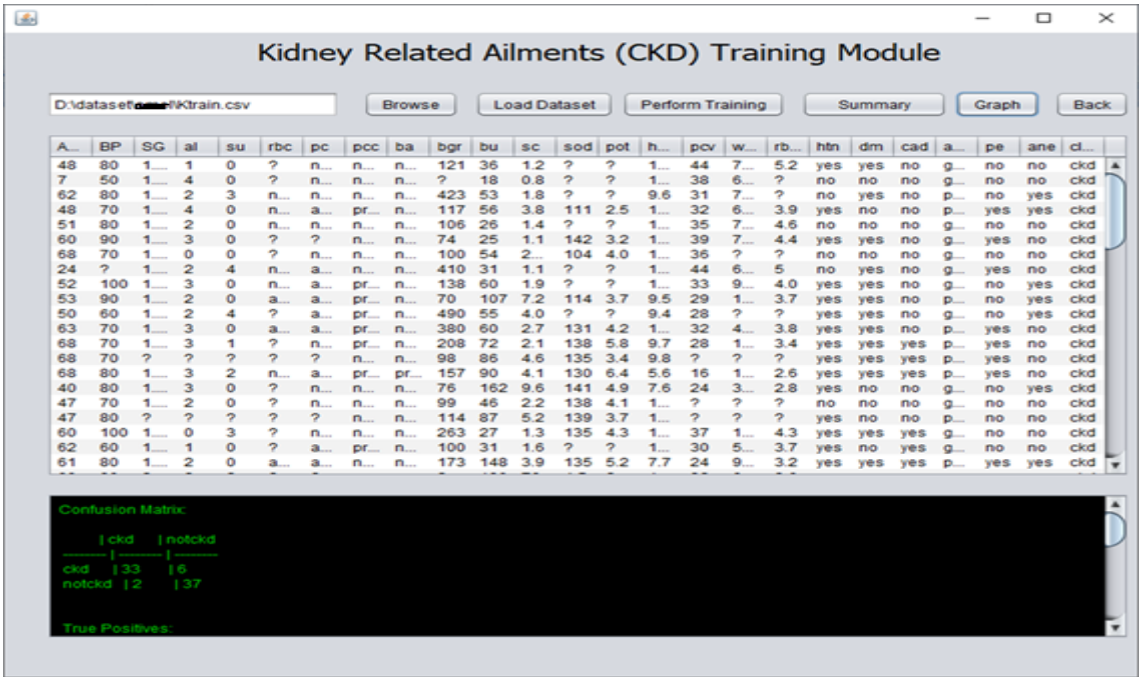


Figure 4.0 Kidney Related Ailments(Training Data)

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

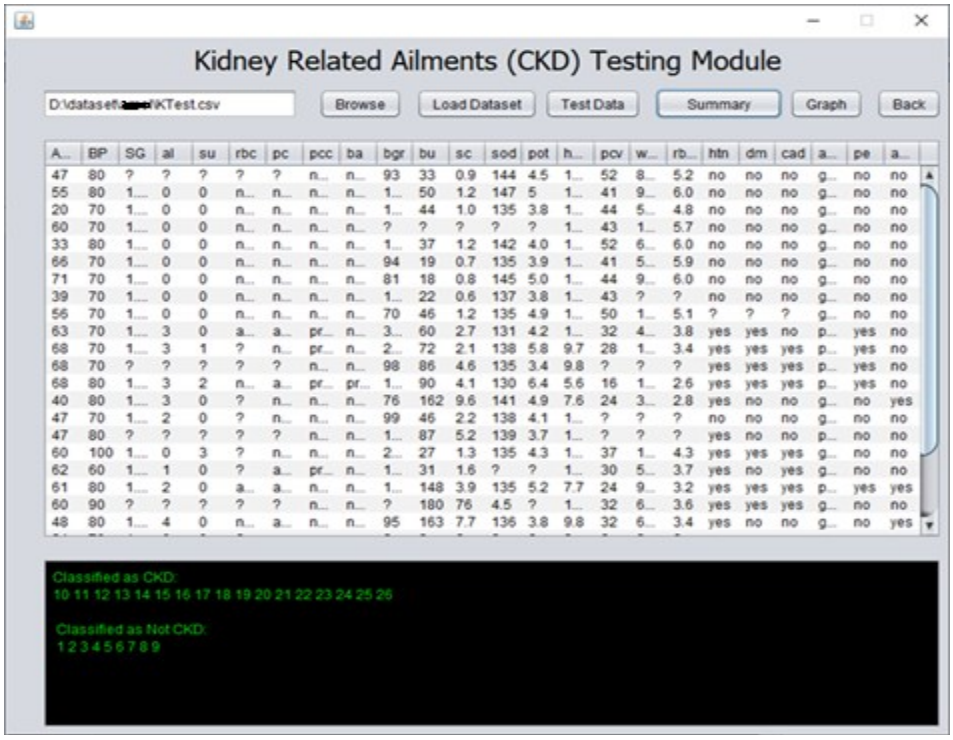
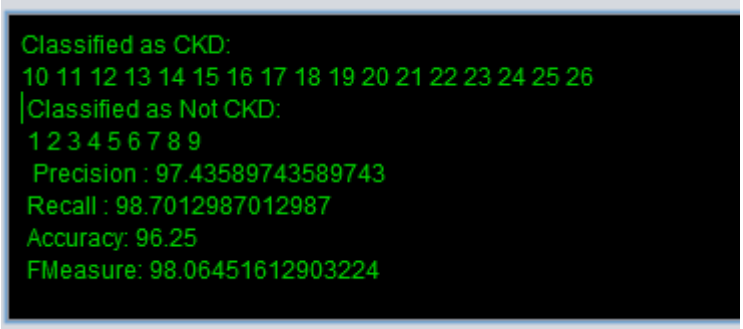


Figure 5.0 Kidney Related Ailments (Testing Data)



iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

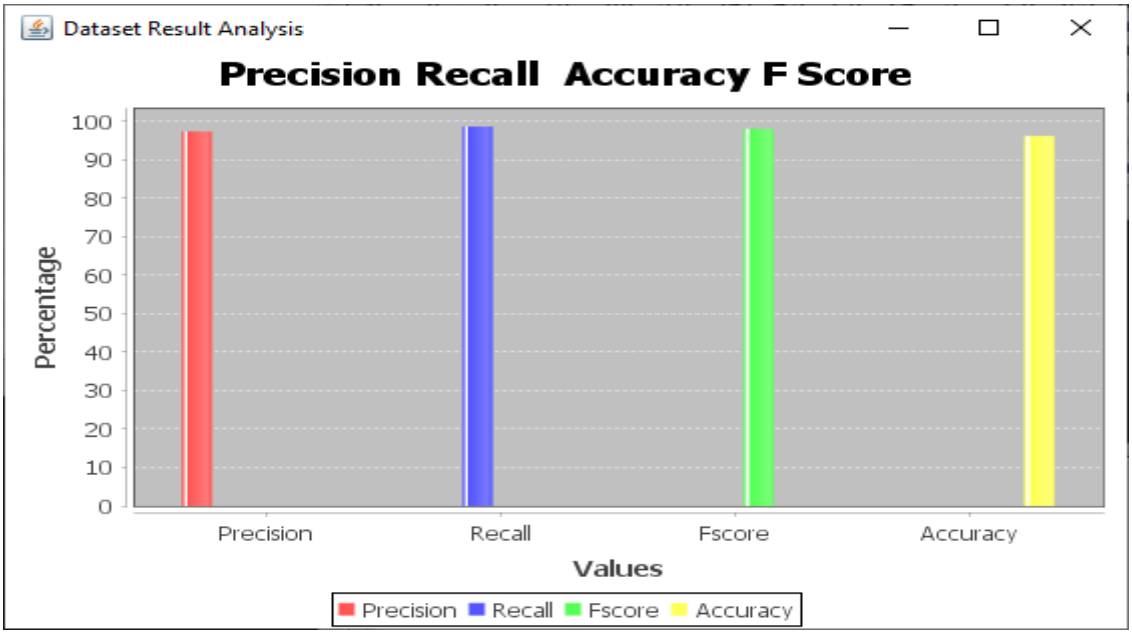


Figure 6.0 Testing and Classification Results

CONCLUSION

This experimental study demonstrates that machine learning techniques can significantly enhance the early prediction of Diabetes Mellitus and Chronic Kidney Disease, two major chronic conditions that require timely medical intervention. The proposed framework systematically integrates data preprocessing, feature selection, model training, and classification to construct an effective disease prediction system. By handling noisy and incomplete healthcare data and selecting the most relevant features, the proposed approach improves both predictive efficiency and classification reliability.

The experimental results show that the proposed method performs consistently well across benchmark datasets, achieving high values of precision, recall, accuracy, F-score, and AUC. In particular, the model attained 96.08% accuracy on the diabetes dataset and 96.25% accuracy on the kidney disease dataset, indicating its capability to support early and dependable diagnosis. Furthermore, the lower average correlation among selected features confirms that the proposed feature selection strategy effectively minimizes redundancy while preserving discriminative information.

Overall, the study confirms that machine learning-driven clinical prediction systems can serve as valuable decision-support tools for healthcare professionals by enabling faster, more accurate, and data-informed diagnosis. The proposed method has strong potential for real-world deployment in hospitals and diagnostic centres, especially in resource-constrained environments where early screening is essential.

As future work, the framework can be extended by incorporating larger real-time clinical datasets, deep learning models, explainable AI techniques, and multimodal patient records to further improve prediction performance and interpretability. Such enhancements would help in building more intelligent, scalable, and patient-centric healthcare prediction systems.

REFERENCES

1) May HT, Anderson JL, Muhlestein JB, Knowlton KU, Horne BD. Intermountain chronic disease risk score (ICHRON) validation for prediction of incident chronic disease diagnoses in an australian primary prevention population. Euro J Intern Med. (2020) 79:81–87. doi: 10.1016/j.ejim.2020.06.009
2) Hegde S, Mundada MR. Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. Int J Pervasive Comput Commun. (2020) 20:145. doi: 10.1108/IJPCC-04-2020-0018
3) Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked. (2019) 16:1–9. doi: 10.1016/j.imu.2019.10020
4) Howard N, Chouikhi N, Adeel A, Dial K, Howard A, Hussain A. BrainOS: a novel artificial brain-alike automatic machine learning framework. Front. Comput. Neurosci. (2020) 14:1–15. doi: 10.3389/fncom.2020. 00016

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

5) Bi X, Zhao X, Huang H, Chen D, Ma Y. Functional brain network classification for Alzheimer's disease detection with deep features and extreme learning machine. Cognit Comput. (2020) 12:513–27. doi: 10.1007/s12559-019-09688-2
6) Guo L. Under The background of healthy china: regulating the analysis of hybrid machine learning in sports activities to control chronic diseases. Measurement. (2020) 164:1–10. doi: 10.1016/j.measurement.2020.107847
7) W.H.O. NonCommunicable Diseases. (2018). Available online at: https://www. who.int/newsroom/fact-sheets/detail/noncommunicable-diseases (accessed December 12, 2021).
8) Hemanth Reddy K, Saranya G. "Prediction of cardiovascular diseases in diabetic patients using machine learning techniques," in Artificial Intelligence Techniques for Advanced Computing Applications, (New York, NY: Springer), p. 299–305 (2020).
9) W.H.O. Cardiovascular diseases (CVDs). (2016). Available online at: https:// www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases- (cvds) (accessed December 12, 2021).
10) Diabetes - A Major Risk Factor for Kidney Disease. National Kidney Foundation. (2020). Available online at: https://www.kidney.org/atoz/content/ diabetes (accessed December 12, 2021).
11) Le TM, Vo TM, Pham TN, Dao SV A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Access 2020;9:7869–84.
12) Julius AO, Ayokunle AO, Ibrahim FO Early diabetic risk prediction using machine learning classification techniques Available from:https://ijisrt.com/early-diabetic-risk-prediction-using-machine-learning-class ification-techniques
13) Shafi S, Ansari GA Early prediction of diabetes disease &classification of algorithms using machine learning approach. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021) Available from:SSRN 3852590 (2021)
14) Khanam JJ, Foo SY A comparison of machine learning algorithms for diabetes prediction. ICT Express 2021;7:432–9.
15) Sisodia D, Sisodia DS Prediction of diabetes using classification algorithms. Procedia Comput Sci 2018;132:1578–85.
16) Agrawal P, Dewangan AK A brief survey on the techniques used for the diagnosis of diabetes-mellitus. Int Res J Eng Tech IRJET 2015;2:1039–43.
17) Rathore A, Chauhan S, Gujral S Detecting and predicting diabetes using supervised learning:An approach towards better healthcare for women. Int J Adv Res Comput Sci 2017;8:1192–4.
18) Hassan AS, Malaserene I, Leema AA Diabetes mellitus prediction using classification techniques. Int J Innov Technol Explor Eng 2020;9:2080–4.
19) Kandhasamy JP, Balamurali S Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci 2015;47:45–51.
20) Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 2013;29:93–9.
21) Nai-Arun N, Moungmai R Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci 2015;69:132–42.
22) Saravananathan K, Velmurugan T Analyzing diabetic data using classification algorithms in data mining. Indian J Sci Technol 2016;9:1–6.
23) Kumari VA, Chitra R Classification of diabetes disease using support vector machine. Int J Eng Res Appl 2013;3:1797–801.
24) Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104–16.
25) Rawat V, Suryakant S A classification system for diabetic patients with machine learning techniques. Int J Math Eng Manag Sci 2019;4:729–44.
26) Perveen S, Shahbaz M, Guergachi A, Keshavjee K Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci 2016;82:115–21.
27) Mujumdar A, Vaidehi V Diabetes prediction using machine learning algorithms. Procedia Comput Sci 2019;165:292–9.
28) Diabetes mellitus affected patients classification and diagnosis through machine learning techniques Procedia Comput Sci 2017;112:2519–28.