**ABSTRACT**

Technological development, including machine learning, has a huge impact on health through an effective analysis of various chronic diseases for more accurate diagnosis and successful treatment. Kidney disease is a major chronic disease associated with aging, hypertension, and diabetes, affecting people 60 and over. Its major cause is the malfunctioning of the kidney in disposing toxins from the blood. This study analyses chronic kidney disease using machine learning techniques based on a chronic kidney disease (CKD) dataset from the UCI machine learning data warehouse. The dataset is pre-processed by completing and normalizing missing data. The most relevant features are selected from the dataset for improved accuracy and reduced training time. The results for selected features of the dataset indicate 99% detection accuracy for CKD. The identified technique is further tested using four patient data samples to predict their CKD.

**INTRODUCTION**

**OVERVIEW**

Kidney disease is considered a major problem for people 60 and above. The major cause is the degeneration of the kidney that reduces the rate of glomerular filtration. This problem, when lasting more than three months, is generally considered as chronic kidney disease. CKD is ranked as the 10th major cause of death in the world. Hypertension, diabetes, and aging are considered leading causes of CKD, in addition to other factors such as high blood pressure, coronary artery disease, and anaemia. If the problem can be detected in early stages, then it is considered feasible to save kidney function for the longer survival of the patient. Early diagnosis of CKD can facilitate its treatment and help avoid costly treatment procedures such as dialysis and transplants.

With machine learning techniques, it is possible to analyse lab records and other information on patients for the early detection of CKD. Low-level data can be transformed into high-level knowledge through the knowledge discovery in databases (KDD). This transformation can help practitioners better understand CKD patterns for its early diagnosis

**PURPOSE**

This study analyses CKD using machine learning techniques using a CKD dataset from the UCI machine learning data warehouse. CKD is detected for 400 instances of chronic kidney patients with 10- fold-cross-validation testing, and the results are compared across a number of classification algorithms. The dataset is pre-processed by completing and normalizing missing data. The most relevant features are selected from the dataset to improve accuracy and reduce training time for machine learning techniques. A set of experiments is conducted using various WEKA-implemented machine learning techniques to detect CMD based on the CKD dataset. The results are compared for detection accuracy across different machine learning techniques.

**LITERATURE SURVEY**

**EXISTING PROBLEM**

CKD is a disease which doesn't shows symptoms at all or in some cases it doesn't show any disease specific symptoms it is hard to predict, detect and prevent such a disease and this could be lead to permanently health damage. Chronic Kidney Disease (CKD) is a condition resulting in insufficient kidney function, where patients have to live with a compromised quality of life. Asia has the highest prevalence of CKD in the world, led by Japan and followed by Taiwan. In Taiwan, CKD has been the eighth leading cause of death since 1997. Compared to other countries, Taiwan has higher incidences and mortality rates, with the prevalence increasing from 1.99% in 1996 to 9.83% in 2003, while awareness about CKD has remained low

CKD is a substantial financial burden on patients, healthcare services, and the government. Treatments of the ESRD with Renal Replacement Therapy are either expensive (haemodialysis and peritoneal dialysis) or complex (transplantation). Taiwan has about 0.1%–0.2% of the population receiving dialysis—contributing to about 7% of the total budget of the National Health Insurance (NHI) program. The association of CKD with other chronic diseases also exacerbates the situation. From the public health perspective, it is therefore imperative to be able to predict the trends in terms of CKD prevalence so that timely decisions can be taken by the decision-makers (ministries, insurers, hospital managers, etc.) to mitigate a potential surge in the number of cases.

Such mitigation measures can include enhanced population screening for CKD-related risks and awareness campaigns, as it has been demonstrated that lifestyle changes (reducing body weight, improving diet, increasing physical activity, reducing alcohol consumption, avoiding smoking, early referral to nephrologists, proper use of medication, and treatments to control other risk factors) are the most effective measures to combat the exacerbation of the condition with minimal associated costs

**PROPOSED SOLUTION**

With the availability of biomedical data, the use of machine-learning techniques in healthcare for developing disease prediction models has become common. Further, methods such as deep learning and techniques like ensemble learning have greatly improved the predictive power of machine learning models. By deriving features from Electronic Health Records (EHR), accurate disease prediction models can be developed. At the patient level, a physician can assess the onset of CKD using laboratory tests by looking at standard parameters such as the glomerular filtration rate (eGFR) and the albumincreatinine ratio. On the other hand, from the public health perspective, laboratory data is typically not available on a large scale. However, two types of data can generally be extracted from the insurance companies’ databases: diagnoses and medications for each patient’s visit at the hospital

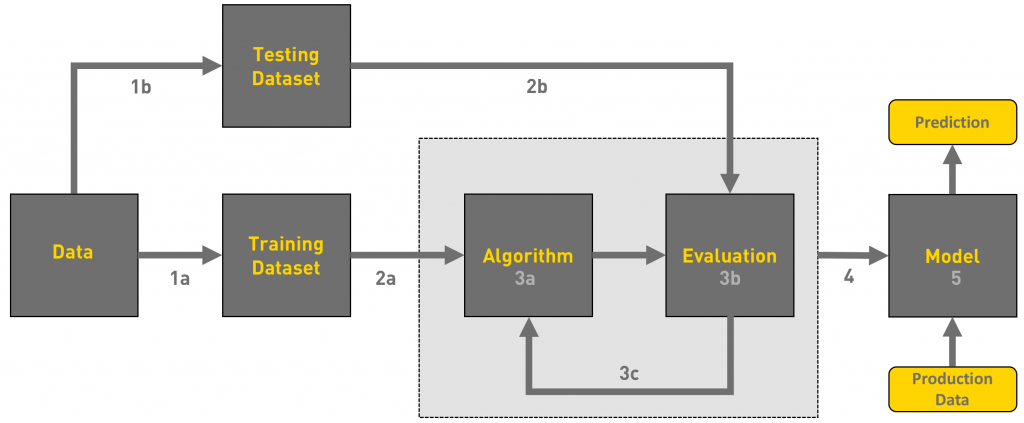
Common approaches for developing disease prediction models with EHR data involve collecting clinical and laboratory data from sources such as billing or claims data, discharge summaries, patient history, etc., and building models on features extracted from them. Some previous studies attempted to use longitudinal data to capture temporal patterns to develop disease prediction models for CKD. Ren et al. (2019) developed a predictive model for kidney disease among patients with hypertension from EHR consisting of textual and numeric information. They proposed a neural network framework based on Bidirectional long short-term memory and auto-encoders to encode the textual and numerical information, respectively. They performed under-sampling to balance the data. They achieved 89.7% accuracy with 10-fold cross-validation

Similarly, some studies used non-temporal EHR data to develop disease prediction models. Song et al. (2019) extracted several significant clinical features from EHR data using an ensemble feature selection method to predict the risk of kidney disease among diabetes patients. They achieved an AUROC of 0.71 on an external validation set

In this paper, we aimed to develop machine-learning models that predict the onset of CKD. The model is based on the insurance claims data (age, sex, comorbidities, and medication). Further, we aim to assess the reliability of the models by identifying the comorbidities and medications that impact the development of CKD the most.

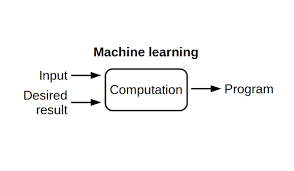
**THEORETICAL ANALYSIS**

**BLOCK DIAGRAM**

****

**Fig** This is a block diagram of how a machine learning model works

**SOFTWARE DEVELOPMENT DESIGNING**



**EXPERIMENTAL INVESTIGATION**

The proposed framework for developing prediction machine learning models and their comparison. The main objective of the present research is to propose a machine learning technique to predict CKD using associative and classification algorithms. The proposed technique generates classification association rules (CARs) to determine techniques with a high percentage of correctly classified instances, and identified classifiers can facilitate early CKD diagnosis. A comparative analysis of the proposed technique is performed using other state-of-the-art techniques. Now we will see details of various stages:

**DATASET**

A dataset with a total of 400 instances with 16 selected attributes is used. The dataset is obtained from the Kaggle website. The attribute “class” is a measurable field with the value “ckd” and indicates an individual with CKD, and “nonckd” indicates an individual with no CKD.

|  |  |
| --- | --- |
| Attributes | Description |
| Age | Range [2 -90] In the year |
| Blood pressure | Range [50 - 180] In mm Hg |
| Red Blood Cell | having two nominal value “normal” or “abnormal” |
| Pus Cell | having two nominal value “normal” or “abnormal” |
| Bacteria | having two nominal value Bacteria is “present” and “not present” |
| Serum Creatinine | numerical value in mgs/dl |
| Haemoglobin | The numerical value in gms |
| Hypertension | having two nominal value “yes” and “no” |
| Diabetes Mellitus | having two nominal value “yes” and “no” |
| Coronary Artery Disease | having two nominal value “yes” and “no” |
| Appetite | having two nominal value Appetite is “good” and “poor” |
| Pedal Edema | having two nominal value Pedal Edema is “yes” and “no” |
| Anaemia | having two nominal value Pedal Edema is “yes” and “no” |
| Class | having the class value “ckd” represent Chronic Kidney Disease and “nonckd” represent Chronic Kidney Disease not present |

***TABLE – DESCRIPTION OF ATTRIBUTES***

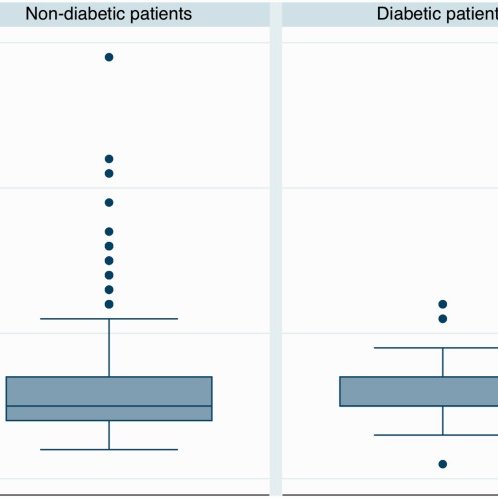
**DATA PREPROCESSING:**

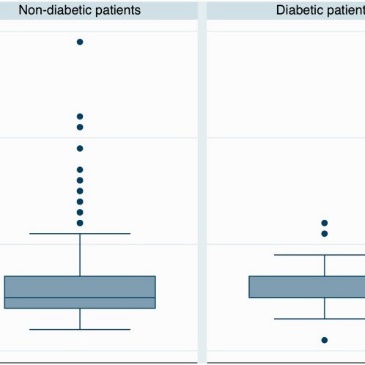
Today’s real-world datasets are susceptible to missing, noisy, redundant, and inconsistent data, especially clinical datasets. Working with low-quality data leads to low-quality results. Therefore, the first step in every machine learning application is to explore the dataset and understand its characteristics in order to make it ready for the modelling stage. This process is commonly known as data pre-processing.

Outliers are extreme values located far away from the feature central tendency. Invalid outliers occur due to data entry errors, which are referred to as a noise in the data. Medical data cannot be treated as other data in dealing with outliers since these outliers could be legitimate (valid) or important. For this reason, each outlier detected in the CKD dataset is checked to know if it is realistic or not.

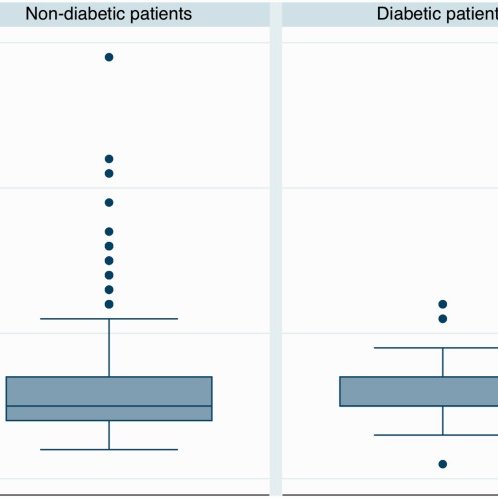
In this study, the extreme data points that go beyond the acceptable range medically have been treated as missing data and then modified as will be described in the missing data section. Box plots have been used to detect outliers in the CKD dataset. There are some outliers detected for blood glucose random that reached 500 mg/dl. However, as mentioned in, the highest blood glucose level recorded in 2008 for a surviving patient reached 2,656 mg/dl. So, these outliers are legitimate and we should not change them

In contrast, for potassium and sodium, three extreme data points are unacceptable. The highest potassium level observed was 7.6 mEq/L. This means that a potassium level with 39 and 47, 2 is impossible and usually due to a mistake. Similarly, with sodium one extreme data point was detected, which is 4.5. Normally, sodium level should be between 135 and 145 mEq/L, and if it is less than 135, then the patient suffers from hyponatremia. For this reason, a value of 4.5 is unacceptable or impossible.

 **fig** Box Plot for blood glucose random

**fig** Box plot for Potassium



**fig** Box plot for Sodium

In real-world datasets, missing data is a very common issue, especially in the medical area. Usually, every patient record and every attribute contains some missing values. However, the chronic kidney disease dataset has 96% of its variables having missing values; 60.75% (243) cases have at least one missing value, and 10% of all values are missing. There are different percentages of missing values for each variable, starting from 0.3% and reaching 38%.

Researchers used single imputation, such as mean and median, to impute the CKD dataset. However, according to Little’s test the missing values in CKD dataset are not missing completely at random. Therefore, single imputation cannot be used for handling missing values.

Data reduction means to reduce the number of features while maintaining a good analytical result. For this purpose, feature selection and features associations or correlation have been studied to remove redundant information.

There is a strong relationship between packed cell volume and haemoglobin and between haemoglobin and red cell count with the correlation coefficient of 0.89 and 0.79 respectively.

Another positive relationship was detected with a correlation coefficient of 0.68 between blood urea and serum creatinine

The process of selecting the most discriminating features in a given dataset is known as feature selection. This process is enhancing the model’s performance, reducing overfitting, and reducing the cost of building a model. Filter feature selection methods selects features that have a stronger relationship with the outcome variable independent to the learning model. Therefore, use a measure or test independent to the learning algorithm to assess a subset of features. In this study, mutual information measure has been used as a feature selection method. Mutual information measures the dependence of any kind of relationships between random variables.

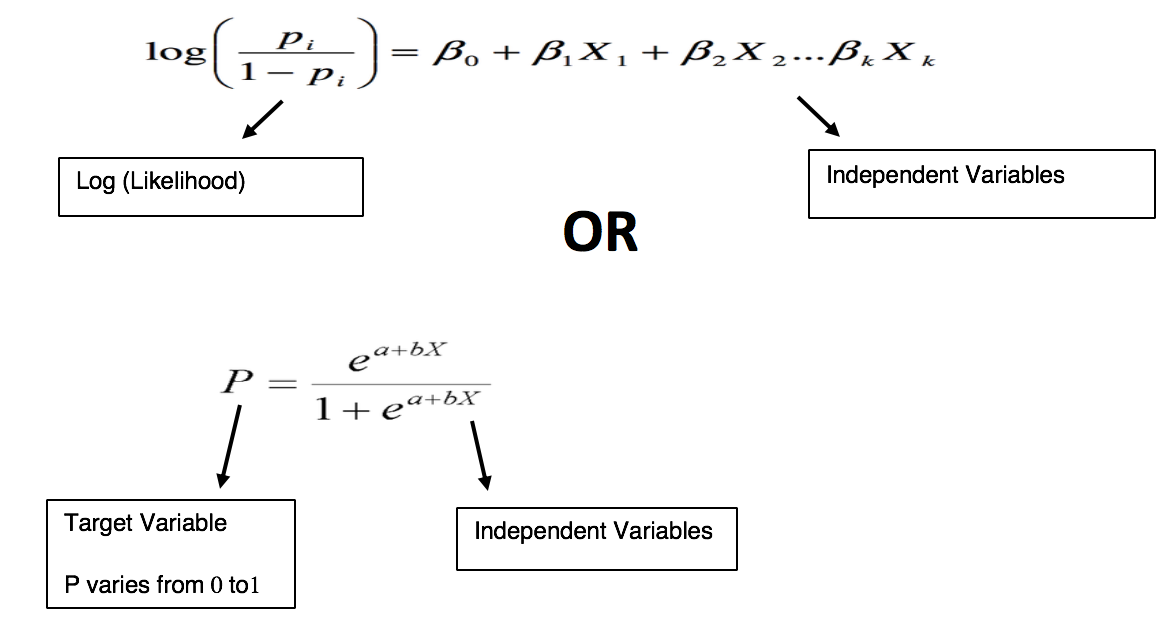
In data transformation, data is transformed into appropriate forms for mining purposes. Data transformation includes normalization, which is the process of scaling the attributes’ values to fall within a small specific range. It is usually applied before feature selection and modeling stages because different scales of attributes complicate the comparison of attributes and influence the ability of algorithms to learn. However, in this study min-max normalization has been applied on numeric data types. Another data transformation has been done on categorical variables. This is because some ML algorithms cannot handle categorical variables, especially in regression problems.

**CKD PREDICTION USING MACHINE LEARNING MODELS**

The paper tries to propose a data mining framework for knowledge discovery on the CKD datasets. Large amounts of CKD datasets are collected. Data preparation and pre-processing is done using the traditional methods of data mining process.

In the modelling stage, four machine learning algorithms have been applied to the dataset to assess their ability to detect CKD. These algorithms are logistic regression (LR), support vector machines (SVM), random forest (RF), and gradient boosting (GB).

1. Logistic regression: Logistic regression, also called logit model or logistic model, is a widely used model to analyze the relationship between multiple independent variables and one categorical dependent variable with the equation of the form:



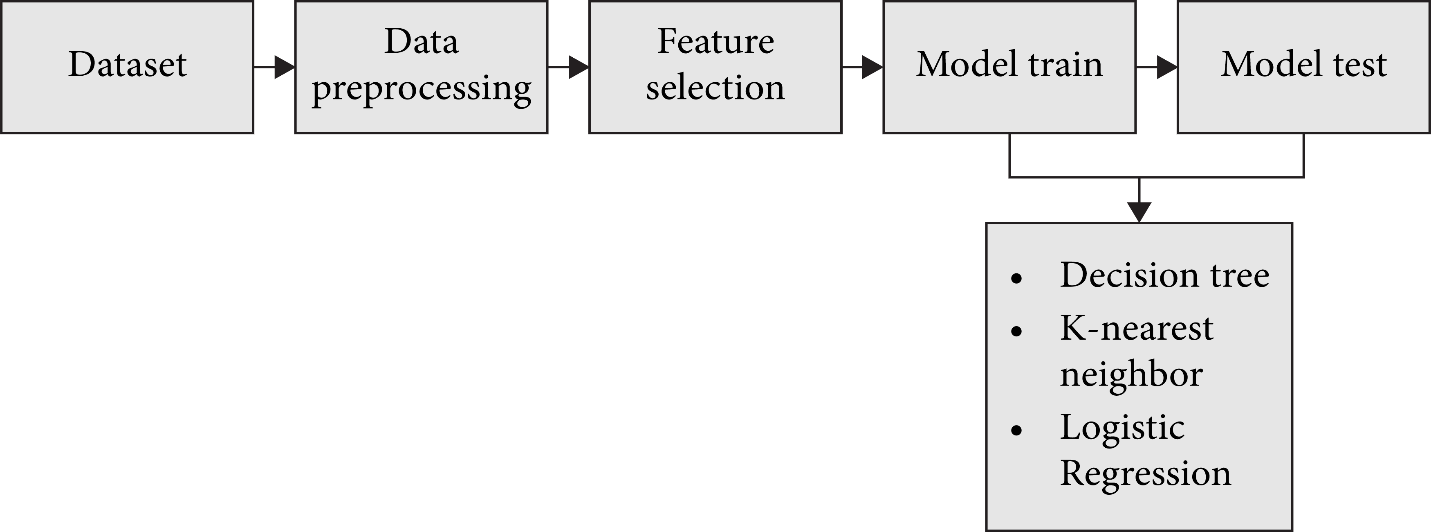
Where is the probability of interest outcome, is an intercept, ,……, are coefficients associated with each variable , and are the values of the predictor variables.

1. Ensemble method: Ensemble method [30] is a strategy for improving predictor or classifier accuracy. Ensemble method uses a combination of models to create an improved composite model to improve the performance. The main idea behind the ensemble technique is to group multiple “weak learners” to come up with a “strong learner”. Two popular techniques for constructing ensembles are bagging and boosting. Both boosting and bagging can be used for prediction as well as classification. Bagging is an ensemble technique where many independent predictors or learners are built and their results are combined using the majority vote, whereas in boosting, the predictors or learners are made sequentially not independently.

Random forest algorithm is an example of the “bagging” technique, whereas the gradient boosting algorithm is an example of the “boosting” technique.

1. Random Forest: Random Forest (RF) is a bagging ensemble approach proposed by Breiman that based on a machine learning mechanism called “decision tree”. In a random forest, the “weak learners” in ensemble terms are decision trees. Random forest imposes the diversity of each tree separately by selecting a random feature. After generating a large number of trees, they vote for the most common class. The random forest algorithm can deal with unbalanced data, it is robust against overfitting, and its runtimes are quite a bit faster.
2. b) Gradient Boosting: Gradient boosting (GB) is an ensemble boosting technique that starts with “regression tree” as “weak learners”. In general, the GB model adds an additive model to minimize the loss function by using a stage-wise sampling strategy. The loss function measures the amount at which the expected value deviates from the real value. Stagewise fashion put more emphasis on samples that are difficult to predict or misclassified. Unlike random forest, in GB, samples that are misclassified have a higher chance of being selected in training data

**FLOW CHART**



**Algorithm:**

Input: Chronic Kidney Disease Dataset

Output: High Accuracy prediction Framework

Step1: Input data

Step2: Pre-process the data

Step 2.1: Convert Categorical values to numerical values

Step 2.2: Replace numerical missing values by Mean

Step2.3: Replace Categorical missing values by Mode

Step3: Construct Classifier Models

Step3.1: Construct the deserved model by testing and training

Step 4: Check the accuracy of the constructed models using confusion matrix.

Step 5: Now create a pkl file to address the model in the flask

Step 6: Create a HTML code for analysis

Step 7: Develop the flask code which links to the HTML web page and create a app.py

Step 8: Now open the link localhost:5000 and predict

**RESULTS AND DISCUSSION**

The result of each classifier has been evaluated using different evaluation metrics and validated against overfitting using 10-fold cross-validation. The nested cross-validation approach also has been applied for the purpose of tuning the models’ parameters. The experiments are conducted using Python 3.3 programming language through the Jupyter Notebook web application

Several libraries from Sciket-learn [35] have been used, which is a free software for the machine learning library in Python. Each model generates different outputs depending on the different values of its parameters.

Each model generates different outputs depending on the different values of its parameters. By using nested crossvalidation, the best performance for LR was with C=1000 and penalty=L2 with an accuracy of 98.9% using F1 measure.

From the evaluation results, all models have an excellent performance against detecting CKD with an accuracy > 97% using haemoglobin, specific gravity, and albumin features. By focusing on specificity and sensitivity, it is seen that all models also have the same specificity of 99.3% except RF (96.6), which means that all models were accurate in identifying the negative or healthy subjects. On the other hand, the highest sensitivity was obtained using the RF algorithm at 99.6%, which represents the percentage of correctly identified CKD patients.

**ADVANTAGES AND DISADVANTAGES**

Early CKD prediction could lead to therapeutic interventions and lifestyle changes, prevention of progression to higher stages, and reduction of dependency as well as costly healthcare spending

However, there remains a need for kidney disease prediction for patients newly diagnosed with T2DM who are at high risk of CKD development

Additionally, early intervention could significantly improve patient quality of life as patients Emerging technologies, pharmacology and therapeutics with CKD report disease and management affecting not only their physical health, but also mental and social health

**CONCLUSION AND FUTURE SCOPE**

However, this research is subject to some limitations related to the dataset used. First, the size of the dataset is considered to be small which may influence the reliability of the results. Second, difficulty finding is another dataset that has the same features in order to compare the results of the datasets.

This study investigates various machine learning techniques, particularly classification and association techniques, to predict CKD. The study analyzes the effects of using feature selection techniques in combination with classification techniques. The results are compared for correctly classified instances, and mean absolute value with and without the feature selection technique. The results note that the best result can be achieved with the Apriori associative algorithm for 97% accuracy.

Future research should analyze different supervised and unsupervised machine learning techniques and feature selection techniques with additional performance metrics for better CKD prediction.

**SOURCE CODE**

https://github.com/smartinternz02/SI-GuidedProject-8286-1642769710