**DIAGNOSING CHRONIC KIDNEY DISEASE**

**USING KNN ALGORITHM**

**ABSTRACT**

Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it induces other diseases. Since there are no conspicuous side effects during the beginning phases of CKD, patients regularly neglect to see the illness. Early discovery of CKD empowers patients to get opportune treatment to enhance the movement of this infection. Machine learning models can successfully help clinicians accomplish this objective because of their quick and precise acknowledgment execution. In this assessment, we propose an KNN and Logistic regression system for diagnosing CKD. The CKD data set was got from the University of California Irvine (UCI) AI store, which has a tremendous number of missing characteristics. KNN attribution was utilized to in the missing qualities, which chooses a few complete examples with the most comparative estimations to handle the missing information for each fragmented example. Missing qualities are generally found, all things considered, clinical circumstances since patients may miss a few estimations for different reasons. After adequately rounding out the fragmented informational index, six AI calculations (strategic relapse, irregular backwoods, uphold vector machine, k-closest neighbour, credulous Bayes classifier and feed forward neural organization) were utilized to set up models. Among these AI models, irregular woodland accomplished the best execution with 99.75% conclusion precision. By breaking down the misjudgements produced by the set up models, we proposed an incorporated model that consolidates calculated relapse and irregular woods by utilizing perceptron, which could accomplish a normal exactness of 99.83% after multiple times of re-enactment. Consequently, we theorized that this philosophy could be appropriate to more confounded clinical information for sickness finding.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. NO.** | **CONTENTS** | | | **PAGE NO.** |
|  | **ABSTRACT** | | | **1** |
|  | **LIST OF TABLES** | | | **vi** |
|  | **LIST OF FIGURES** | | | **vii** |
|  | **LIST OF SYMBOLS AND ABBREVIATIONS** | | | **viii** |
| **1** | **INTRODUCTION** | | | **1** |
|  | 1.1 | | CHRONIC KIDNEY DISEASE | 1 |
|  | 1.2 | | MACHINE LEARNING | 3 |
|  | 1.3 | | KNN IMPUTATION | 4 |
| **2** | **LITERATURE SURVEY** | | | **5** |
|  | 2.1MECHANICAL ANISOTROPY ASSESSMENT IN KIDNEY CORTEX USING ARFI PEAK DISPLACEMENT: PRECLINICAL VALIDATION AND PILOT IN VIVO CLINICAL RESULTS IN KIDNEY ALLOGRAFTS. | | | 5 |
|  | 2.2 IN VIVO DETECTION OF CHRONIC KIDNEY DISEASE USING TISSUE DEFORMATION FIELDS FROM DYNAMIC MR IMAGING. | | | 6 |
|  | 2.3 EXPLAINABLE PREDICTION OF CHRONIC RENAL DISEASE IN THE COLOMBIAN POPULATION USING NEURAL NETWORKS AND CASE-BASED REASONING. | | | 7 |
|  | 2.4 NEURAL NETWORK AND SUPPORT VECTOR MACHINE FOR THE PREDICTION OF CHRONIC KIDNEY DISEASE: A COMPARATIVE STUDY | | | 8 |
|  | **2.5** APPLICATION OF MACHINE LEARNING IN PREDICTING PERFORMANCE FOR COMPUTER ENGINEERING STUDENTS: A CASE STUDY | | | 9 |
|  | 2.6 A MACHINE LEARNING BASED APPROACH TO IDENTIFY PROTECTED HEALTH INFORMATION IN CHINESE CLINICAL TEXT | | | 10 |
|  | **2.7** CLASSIFICATION OF FOETAL DISTRESS AND HYPOXIA USING MACHINE LEARNING APPROACHES | | | 11 |
|  | **2.8** COMPARISON ANALYSIS OF MACHINE LEARNING ALGORITHMS TO RANK ALZHEIMER’S DISEASE RISK FACTORS BY IMPORTANCE | | | 13 |
|  | **2.9** PREDICTING DIABETES MELLITUS WITH MACHINE LEARNING TECHNIQUES | | | 14 |
|  | **2.10** MOMENTUM PRINCIPAL SKEWNESS ANALYSIS | | | 15 |
|  | **2.11** PREDICTING ACUTE KIDNEY INJURY IN CANCER PATIENTS USING HETEROGENEOUS AND IRREGULAR DATA | | | 16 |
|  | **2.12** USING RESISTIN, GLUCOSE, AGE AND BMI TO PREDICT THE PRESENCE OF BREAST CANCER | | | 17 |
|  | **2.13** A NEW EFFECTIVE MACHINE LEARNING FRAMEWORK FOR SEPSIS DIAGNOSIS | | | 18 |
|  | **2.14** EARLY PREDICTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING SUPPORTED BY PREDICTIVE ANALYTIC**S** | | | 19 |
|  | **2.15** COMPARISON OF VARIABLE SELECTION METHODS FOR CLINICAL PREDICTIVE MODELING | | | 20 |
|  | **2.16** DIAGNOSIS OF CHRONIC KIDNEY DISEASE BY USING RANDOM FOREST | | | 21 |
|  | **2.17** DIAGNOSIS OF CHRONIC KIDNEY DISEASE BASED ON SUPPORT VECTOR MACHINE BY FEATURE SELECTION METHODS | | | 22 |
|  | **2.18** CHRONIC KIDNEY DISEASE, BASAL INSULIN GLARGINE AND HEALTH OUTCOMES IN PEOPLE WITH DYSGLYCEMIA: THE ORIGIN STUDY | | | 23 |
|  | **2.19** MACHINE-LEARNING-BASED CLASSIFICATION OF REAL-TIME TISSUE ELASTOGRAPHY FOR HEPATIC FIBROSIS IN PATIENTS WITH CHRONIC HEPATITIS | | | 24 |
|  | **2.20** PERFORMANCE EVALUATION ON MACHINE LEARNING CLASSIFICATION TECHNIQUES FOR DISEASE CLASSIFICATION AND FORECASTING THROUGH DATA ANALYTICS FOR CHRONIC KIDNEY DISEASE (CKD) | | | 26 |
| **3** | MATERIALS AND METHODS | | | 27 |
|  | 3.1 | EXISTING SYSTEM | | 27 |
|  |  | 3.1.1 DISADVANTAGES | | 28 |
|  | 3.2 | PROPOSED SYSTEM | | 29 |
|  |  | 3.2.1 ADVANTAGES | | 30 |
|  | 3.3 | MODULE DESCRIPTION | | 31 |
|  | 3.4 | SYSTEM REQUIREMENTS  3.4.1 Hardware Requirements  3.4.2 Software Requirements | | 38 |
| **4** | **EXPERIMENTAL SETUP AND PROCEDURE** | | | **39** |
| **5** | **RESULTS AND DISCUSSION** | | | **41** |
| **6** | **CONCLUSION** | | | **42** |
|  | **REFERENCES** | | | 43 |

LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| **TABLE NO.** | **TITLE** | **PAGE NO.** |
| 1.1 | Integrated model for Acc,Sen,Spec and F1 represent the accuracy,sensitivity,specifity and F1 score, respectively. | 41 |

LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| 4.1 | Integrated model for Acc ,Sen, Spec and F1 | 35 |

LIST OF SYMBOLS AND ABBREVIATIONS

**SYMBOLS**

|  |  |  |  |
| --- | --- | --- | --- |
| g | - | Acceleration due to gravity | m/s2 |
|  | - | Area | m2 |

**GREEK SYMBOLS**

|  |  |  |  |
| --- | --- | --- | --- |
| ρ | - | Density of air | kg/m3 |
|  | - | Efficiency | % |

**ABBREVIATIONS**

|  |  |  |
| --- | --- | --- |
| CKD | - | Chronic kidney disease |
| ACEI | - | Angiotensin converting enzyme inhibitors |
| KNN | - | K-Nearest Neighbours Algorithm |
| LR | - | Logistic Regression |
| RF | - | Random Forest |
| DT | - | Decision Tree |

INTRODUCTION

Their studies have achieved good results in the diagnosis of CKD. In the above models, the mean imputation is used to ll in the missing values and it depends on the diagnostic categories of the samples. As a result, their method could not be used when the diagnostic results of the samples are unknown. In reality, patients might miss some measurements for various reasons before diagnosing. In addition, for missing values in categorical variables, data obtained using mean imputation might have a large deviation from the actual values. For example, for variables with only two categories, we set the categories to 0 and 1, but the mean of the variables might be between 0 and 1. developed an based on feature selection technology, the proposed models reduced the computational cost through feature selection, and the range of accuracy in those model was from 97.75%-98.5%

**1.1 CHRONIC KIDNEY DISEASE**

Chronic kidney disease (CKD) is a type of kidney disease in which there is gradual loss of kidney function over a period of months to years. Initially there are generally no symptoms; later, symptoms may include leg swelling, feeling tired, vomiting, loss of appetite, and confusion. Complications include an increased risk of heart disease, high blood pressure, bone disease, and anaemia. Causes of chronic kidney disease include diabetes, high blood pressure, glomerulonephritis, and polycystic kidney disease. Risk factors include a family history of chronic kidney disease. Diagnosis is by blood tests to measure the estimated glomerular filtration rate (eGFR), and a urine test to measure albumin .Ultrasound or kidney biopsy may be performed to determine the underlying cause. Several severity-based staging systems are in use. Screening at-risk people is recommended. Initial treatments may include medications to lower blood pressure, blood sugar, and cholesterol. Angiotensin converting enzyme inhibitors (ACEIs) or angiotensin II receptor antagonists (ARBs) are generally first-line agents for blood pressure control, as they slow progression of the kidney disease and the risk of heart disease. Loop diuretics may be used to control edema and, if needed, to further lower blood pressure.NSAIDs should be avoided. Other recommended measures include staying active, and certain dietary changes such as a low-salt diet and the right amount of protein.Treatments for anemia and bone disease may also be required. Severe disease requires hemodialysis, peritoneal dialysis, or a kidney transplant for survival. Blood pressure is increased due to fluid overload and production of vasoactive hormones created by the kidney via the renin–angiotensin system, increasing the risk of developing hypertension and heart failure.

Urea accumulates, leading to azotemia and ultimately uremia (symptoms ranging from lethargy to pericarditis and encephalopathy). Due to its high systemic concentration, urea is excreted in eccrine sweat at high concentrations and crystallizes on skin as the sweat evaporates ("uremic frost").Potassium accumulates in the blood (hyperkalemia with a range of symptoms including malaise and potentially fatal cardiac arrhythmias). Hyperkalaemia usually does not develop until the glomerular filtration rate falls to less than 20–25 ml/min/1.73 m2, at which point the kidneys have decreased ability to excrete potassium. Hyperkalaemia in CKD can be exacerbated by academia (which leads to extracellular shift of potassium) and from lack of insulin. Changes in mineral and bone metabolism that may cause 1) abnormalities of calcium, phosphorus (phosphate), parathyroid hormone, or vitamin D metabolism; 2) abnormalities in bone turnover, mineralization, volume, linear growth, or strength (kidney osteodystrophy); and 3) vascular or other soft-tissue calcification.CKD-mineral and bone disorders have been associated with poor outcomes.Metabolic acidosis may result from decreased capacity to generate enough ammonia from the cells of the proximal tubule.[20] Acidemia affects the function of enzymes and increases excitability of cardiac and neuronal membranes by the promotion of hyperkalemia.Anemia is common and is especially prevalent in those requiring haemodialysis. It is multifactoral in cause, but includes increased inflammation, reduction in erythropoietin, and hyperuricemia leading to bone marrow suppression.In later stages, cachexia may develop, leading to unintentional weight loss, muscle wasting, weakness and anorexia.

**1.2 MACHINE LEARNING**

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks.

For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step. The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. For example, to train a system for the task of digital character recognition, the MNIST dataset of handwritten digits has often been used.

**1.3 KNN IMPUTATION**

KNN Imputer by scikit-learn is a widely used method to impute missing values. It is widely being observed as a replacement for traditional imputation techniques. In today’s world, data is being collected from a number of sources and is used for analysing, generating insights, validating theories, and whatnot. This data collected from different resources may often have some information missing. This may be due to a problem in the data collection or extraction process that could be a human error. Dealing with these missing values, thus becomes an important step in data pre-processing. The choice of method of imputation is crucial since it can significantly impact one’s work. A handful of literature in statistics deals with the source of missing values and ways to overcome the issue. The best way is to impute these missing observations with an estimated value. In this article, we introduce a guide to impute missing values in a dataset using values of observations for neighbouring data points. For this, we use the very popular KNN Imputer by scikit-learn k-Nearest Neighbours Algorithm. Missing values in a dataset can be a hornet’s nest for any data scientist. Variables with missing values can be a non-trivial problem as there is no easy way out to deal with them. Generally, if the proportion of missing observations in data is small relative to the total number of observations, we can simply remove those observations. However, this is not the most often case. Deleting the rows containing missing values may lead to parting away with useful information or patterns.

This happens when the missing values have no hidden dependency on any other variable or any characteristic of observations. If a doctor forgets to record the age of every tenth patient entering an ICU, the presence of missing value would not depend on the characteristic of the patients. n this case, the probability of missing value depends on the characteristics of observable data. In survey data, high-income respondents are less likely to inform the researcher about the number of properties owned. The missing value for the variable number of properties owned will depend on the income variable. This happens when the missing values depend on both characteristics of the data and also on missing values. In this case, determining the mechanism of the generation of missing value is difficult. For example, missing values for a variable like blood pressure may partially depend on the values of blood pressure as patients who have low blood pressure are less likely to get their blood pressure checked at frequently.

**LITERATURE REVIEW**

**2.1 MECHANICAL ANISOTROPY ASSESSMENT IN KIDNEY CORTEX USING ARFI PEAK DISPLACEMENT: PRECLINICAL VALIDATION AND PILOT IN VIVO CLINICAL RESULTS IN KIDNEY ALLOGRAFTS**

**Md Murad Hossain, et al.,** has proposed in this work kidney is an anisotropic organ, with higher elasticity along versus across nephrons. The degree of mechanical anisotropy in the kidney may be diagnostically relevant if properly exploited; however, if improperly controlled, anisotropy may confound stiffness measurements. The purpose of this study is to demonstrate the clinical feasibility of Acoustic Radiation Force (ARF) induced peak displacement (PD) measures for both exploiting and obviating mechanical anisotropy in the cortex of human kidney allografts, in vivo. Validation of the imaging methods is provided by pre-clinical studies in pig kidneys, in which ARF-induced PD values were statistically significantly higher (p0.01). Similar results were demonstrated in vivo in the kidney allografts of 14 patients. The symmetric ARF produced PD measures with no statistically significant difference (p>0.01) between along versus across alignments, but the asymmetric ARF yielded PD ratios that remained constant over a six-month observation period posttransplantation, consistent with stable serum creatinine level and urine protein to creatinine ratio in the same patient population (p>0.01). The results of this pilot in vivo clinical study suggest the feasibility of: 1) implementing symmetrical ARF to obviate mechanical anisotropy in the kidney cortex when anisotropy is a confounding factor, and 2) implementing asymmetric ARF to exploit mechanical anisotropy when mechanical anisotropy is a potentially relevant biomarker.

This study demonstrates the clinical feasibility of ARFIPD ratio as a measure of mechanical anisotropy in the allografts of renal transplant patients. In healthy allografts, PD ratios were stable over two, four, and six months posttransplantations. Preclinical validation studies in pig kidneys showed that PD-ratios linearly correlated with SWEI-derived shear moduli ratios ex vivo as well as in vivo at baseline, with arterial ligation, and with venous ligation. Further, mechanical anisotropy was exploited by using an asymmetrical ARF, but a symmetric ARF generally obviated anisotropy and enabled angle independent mechanical property assessment in the anisotropic cortex. Thus, when mechanical anisotropy is of interest as a renal biomarker, an asymmetric ARF should be implemented. However, when anisotropy is considered a confounding factor due to unknown or uncontrollable transducer orientation with respect to nephron alignment, a symmetric ARF should be used. This work represents a critical first step towards clinically translating ARFI PD ratio as a relevant, new, noninvasive biomarker for monitoring kidney transplant health **[1].**

**2.2 IN VIVO DETECTION OF CHRONIC KIDNEY DISEASE USING TISSUE DEFORMATION FIELDS FROM DYNAMIC MR IMAGING**

**Erlend Hodneland, Eirik Keilegavlen et al.,** has proposed in this work Chronic kidney disease is a serious medical condition characterized by gradual loss in kidney function. Early detection and diagnosis is mandatory for prognostic improvement. Hence, in the current work we explore the use of image registration methods for detecting pathological changes in patients with chronic kidney disease. Methods: Ten healthy volunteers and nine patients with presumed chronic kidney disease underwent dynamic T1 weighted imaging without contrast agent. From real and simulated dynamic time series, kidney deformation fields were estimated using a poroelastic deformation model. From the deformation fields several quantitative parameters reflecting pressure gradients, and volumetric and shear deformations were computed. Eight of the patients also underwent biopsy as a gold standard. Results: We found that the absolute deformation, normalized volume changes, as well as pressure gradients correlated significantly with arteriosclerosis from biopsy assessments. Furthermore, our results indicate that current image registration methodologies are lacking sensitivity to recover mild changes in tissue stiffness. Conclusion: Image registration applied to dynamic time series should be further explored as a tool for invasive measurements of arteriosclerosis.

Significance: Under the assumption that the proposed framework can be further developed in terms of sensitivity and specificity, it can provide clinicians with a non-invasive tool of high spatial coverage available for characterization of arteriosclerosis and potentially other pathological changes observed in chronic kidney disease. In the current pilot study we propose a framework for detecting pathological changes in stiffness using image registration. First results on a small cohort of patients indicate that the absolute deformation, normalized volume changes, as well as pressure gradients demonstrate a statistical correlation to arteriosclerosis in the kidneys, and hence could be used as a proxy for arteriosclerosis grade. The simulation study indicated that image registration has a borderline sensitivity close to the stiffness changes seen in mild CKD. Still, the algorithms were capable of revealing significant correlations with biopsy assessments, and future implementations will demonstrate whether image registration has sufficient abilities to be used in personalized medicine as a complementary tool for quantification of mild pathological changes in tissue stiffness **[2].**

**2.3 EXPLAINABLE PREDICTION OF CHRONIC RENAL DISEASE IN THE COLOMBIAN POPULATION USING NEURAL NETWORKS AND CASE-BASED REASONING**

**Gabriel R. Vásquez-Morales , Sergio M. Martínez-Monterrubio et al.,** has proposed in this work presents a neural network-based classifier to predict whether a person is at risk of developing chronic kidney disease (CKD). The model is trained with the demographic data and medical care information of two population groups: on the one hand, people diagnosed with CKD in Colombia during 2018, and on the other, a sample of people without a diagnosis of this disease. Once the model is trained and evaluation metrics for classification algorithms are applied, the model achieves 95% accuracy in the test data set, making its application for disease prognosis feasible. However, despite the demonstrated efficiency of the neural networks to predict CKD, this machine-learning paradigm is opaque to the expert regarding the explanation of the outcome. Current research on eXplainable AI proposes the use of twin systems, where a black-box machine-learning method is complemented by another white-box method that provides explanations about the predicted values. Case-Based Reasoning (CBR) has proved to be an ideal complement as this paradigm is able to find explanatory cases for an explanation-by-example justification of a neural network’s prediction. In this work, we apply and validate a NN-CBR twin system for the explanation of CKD predictions. As a result of this research, 3,494,516 people were identified as being at risk of developing CKD in Colombia, or 7% of the total population.

All previous studies for CKD prediction use laboratory test information as input variables for model training and a relatively small pool of patients is available. However, in this work we have collected a training data set corresponding to 40,000 people, and for the implementation of the model the information from health care performed on 39,277,086 people during the years 2009 to 2018 was used. These include demographic information, such as sex, age, ethnicity and place of residence, as well as the history of the pathologies that have been diagnosed to the person, coded according to the International Classification of Diseases (ICD). From this novel dataset, we have developed a neural network approach that can predict the risk of developing chronic kidney disease with an accuracy of 95%. This is a remarkable result as the most similar study achieves an accuracy of 89.7% [31],. This result was obtained by training a neuronal network with 5 layers: an input layer with 7,492 neurons, corresponding to the variables or characteristics of the model, 3 hidden layers with 500, 100 and 50 neurons respectively, and an output layer with a single neuron representing the class of the binary classification problem. For network training, a ReLU activation function was used in the hidden layers and a sigmoid function to find the probability in the output layer. Adam was used as the training algorithm of the model, which showed a higher convergence speed compared to other more traditional algorithms such as gradient descent and stochastic gradient descent (SGD). To combat the effect of overfitting in the network, regularization was used using the dropout technique, in conjunction with early stopping **[3].**

**2.4 NEURAL NETWORK AND SUPPORT VECTOR MACHINE FOR THE PREDICTION OF CHRONIC KIDNEY DISEASE: A COMPARATIVE STUDY**

**Njoud Abdullah Almansour, Hajra Fahim Syed et al.,** has proposed in this work aims to assist in the prevention of Chronic Kidney Disease (CKD) by utilizing machine learning techniques to diagnose CKD at an early stage. Kidney diseases are disorders that disrupt the normal function of the kidney. As the percentage of patients affected by CKD is significantly increasing, effective prediction procedures should be considered. In this work, we focus on applying different machine learning classification algorithms to a dataset of 400 patients and 24 attributes related to diagnosis of chronic kidney disease. The classification techniques used in this study include Artificial Neural Network (ANN) and Support Vector Machine (SVM). To perform experiments, all missing values in the dataset were replaced by the mean of the corresponding attributes. Then, the optimized parameters for the Artificial Neural Network (ANN) and Support Vector Machine (SVM) techniques were determined by tuning the parameters and performing several experiments. The final models of the two proposed techniques were developed using the best-obtained parameters and features.

The empirical results from the experiments indicated that ANN performed better than SVM, with accuracies of 99.75% and 97.75%, respectively, indicating that the outcome of this study is very promising. Support Vector Machine (SVM) and Artificial Neural Network (ANN) are two of the most known machine learning techniques. Both techniques have advantages and have been proven to perform excellently in several fields. ANN has been proposed as a new model to better predict CKD and has been compared against SVM, which has exhibited the highest accuracy in previous studies. The dataset was first preprocessed, and the missing values were replaced. A 10-fold crossvalidation was used as the algorithm for the partitioning of the training and test datasets with the ratio (90: 10). The optimized parameters for Support Vector Machine and Artificial Neural Network were identified. Several experiments were performed using different values of the parameters for both techniques. It was found that ANN performs better with an accuracy of 99.75% using the optimized features, while SVM had a performance accuracy of 97.75% **[4].**

**2.5 APPLICATION OF MACHINE LEARNING IN PREDICTING PERFORMANCE FOR COMPUTER ENGINEERING STUDENTS: A CASE STUDY**

**Diego Buenaño-Fernández , David Gil et al.,** has proposed in this work present work proposes the application of machine learning techniques to predict the final grades (FGs) of students based on their historical performance of grades. The proposal was applied to the historical academic information available for students enrolled in the computer engineering degree at an Ecuadorian university. One of the aims of the university’s strategic plan is the development of a quality education that is intimately linked with sustainable development goals (SDGs). The application of technology in teaching–learning processes (Technology-enhanced learning) must become a key element to achieve the objective of academic quality and, as a consequence, enhance or benefit the common good. Today, both virtual and face-to-face educational models promote the application of information and communication technologies (ICT) in both teaching–learning processes and academic management processes. This implementation has generated an overload of data that needs to be processed properly in order to transform it into valuable information useful for all those involved in the field of education. Predicting a student’s performance from their historical grades is one of the most popular applications of educational data mining and, therefore, it has become a valuable source of information that has been used for different purposes.

Nevertheless, several studies related to the prediction of academic grades have been developed exclusively for the benefit of teachers and educational administrators. Little or nothing has been done to show the results of the prediction of the grades to the students. Consequently, there is very little research related to solutions that help students make decisions based on their own historical grades. This work proposes a methodology in which the process of data collection and pre-processing is initially carried out, and then in a second stage, the grouping of students with similar patterns of academic performance was carried out. In the next phase, based on the identified patterns, the most appropriate supervised learning algorithm was selected, and then the experimental process was carried out. Finally, the results were presented and analyzed. The results showed the effectiveness of machine learning techniques to predict the performance of students. After we have verified the model proposed, the most imminent future work is to analyze and design a big data architecture that supports the processing of the large amount of academic data that the university generates periodically. This academic data should be also complemented with other data, such as personal and socio-economic information of the student and information on the student learning assessment system, among others. This large volume of data can be increased by scaling up the proposal of this work for all the university’s degrees. To define the project architecture, it is not recommended to use a traditional approach based on a data warehouse; rather, due to the nature of the proposed project, it will be necessary to create a documented, scalable, and flexible database that can support large indexing and data consultation by students, teachers, and educational administrators. Therefore, we plan to design an architecture that uses big data tools, such as Hadoop and MongoDB, in parallel **[5].**

**2.6 A MACHINE LEARNING BASED APPROACH TO IDENTIFY PROTECTED HEALTH INFORMATION IN CHINESE CLINICAL TEXT**

**Liting Dua , Chenxi Xia et al.,** has proposed in this work with the increasing application of electronic health records (EHRs) in the world, protecting private information in clinical text has drawn extensive attention from healthcare providers to researchers. De-identification, the process of identifying and removing protected health information (PHI) from clinical text, has been central to the discourse on medical privacy since 2006. While de-identification is becoming the global norm for handling medical records, there is a paucity of studies on its application on Chinese clinical text. Without efficient and effective privacy protection algorithms in place, the use of indispensable clinical information would be confined. This study presented a machine learning based approach to identify protected health information in Chinese clinical text and combined it with post-processing rules to improve the recall of the less frequent PHI types. We constructed a Chinese clinical text corpus with PHI tags by manual annotation. The descriptive statistics of PHI in our corpus have shown the frequent and diverse distribution of PHI in Chinese clinical text, which proved the necessity and urgency for the furtherance of Chinese clinical text de-identification researches.

Finally, the high-performance of our model demonstrates that based on the common features, the CRF machine learning algorithm widely used in English clinical text de-identification is applicable for Chinese clinical text de-identification. The descriptive statistics of our corpus have proven the wide range and diversity of the distribution of PHI entities in Chinese clinical text. This confirms the necessity and urgency for a de-identification tool to parse PHI from Chinese clinical text to protect patients’ privacy during the process of communicating clinical information with sensitive segments. Regarding the 18 categories of PHI that must be removed from the EHRs defined by HIPAA, China does not have the same degree of legislation in place to protect patient privacy and regulate the storage and sharing of such clinical information. Moreover, there is no relevant research to define and analyze PHI categories in Chinese clinical text. In Jian and colleague’s work, only three PHI categories of name, address, and hospital were annotated in their corpus [27]. In our study, the statistics on the actual distribution of PHI in Chinese clinical text may warrant further research on clinical text de-identification and the pursuant legislation in China **[6].**

**2.7 CLASSIFICATION OF FOETAL DISTRESS AND HYPOXIA USING MACHINE LEARNING APPROACHES**

**Rounaq Abbas1 , Abir Jaafar Hussain et al.,** has proposed in this work Foetal distress and hypoxia (oxygen deprivation) is considered as a serious condition and one of the main factors for caesarean section in the obstetrics and Gynecology department. It is the third most common cause of death in new-born babies. Many foetuses that experienced some sort of hypoxic effects can develop series risks including damage to the cells of the central nervous system that may lead to life-long disability (cerebral palsy) or even death. Continuous labour monitoring is essential to observe the foetal well being. Foetal surveillance by monitoring the foetal heart rate with a cardiotocography is widely used. Despite the indication of normal results, these results are not reassuring, and a small proportion of these foetuses are actually hypoxic. In this work, machine-learning algorithms are utilized to classify foetuses which are experiencing oxygen deprivation using PH value (a measure of hydrogen ion concentration of blood used to specify the acidity or alkalinity) and Base Deficit of extra cellular fluid level (a measure of the total concentration of blood buffer base that indicates the metabolic acidosis or compensated respiratory alkalosis) as indicators of respiratory and metabolic acidosis, respectively, using open source partum clinical data obtained from Physionet. Six well know machine learning classifier models are utilised in our experiments for the evaluation; each model was presented with a set of selected features derived from the clinical data.

Classifier’s evaluation is performed using the receiver operating characteristic curve analysis, area under the curve plots, as well as the confusion matrix. Our simulation results indicate that machine-learning algorithms provide viable methods that could delivery improvements over conventional analysis. In this work, we have presented various machine learning approaches for the classification of foetal hypoxia. We reviewed related works and problem parameters including the main characteristics of foetal distress and the essential nature of early prediction. In our study, we have investigated the application of machine learning for foetal hypoxia classification. Experimental data was obtained from open source Physionet dataset. We applied a series of test models to extracted data, whose performances were evaluated using both graphical forms of analysis, including the ROC plot, and scalar summary indices including sensitivity, specificity, Kappa, and overall accuracy. A classification scheme is proposed in which the foetal experience of hypoxia is confirmed by a blood gas and acid-base assessment with evidence of a significant respiratory (PH value) and metabolic acidosis (BDecf value). Although some studies show that resampling of imbalanced data may lower the model performance, in our study results remained promising. Statistical and graphical comparison between various classification methods has shown that RF and NN are the strongest classifiers within our trials of foetal hypoxia detection, and the most models used for such classification compared to other classifiers **[7].**

**2.8 COMPARISON ANALYSIS OF MACHINE LEARNING ALGORITHMS TO RANK ALZHEIMER’S DISEASE RISK FACTORS BY IMPORTANCE**

**Mohamed Mahyoub, Dr Martin Randles et al.,** has proposed in this work people have always feared aging, and the increasing rate of dementia disease caused this fear to twofold. Dementia is irreversible, unstoppable and has no known cure. According to Alzheimer's Disease International 2015 and World Alzheimer Report 2015, the estimated financial cost for healthcare services of Alzheimer's Disease is $1 Trillion in 2018. This work discusses the importance of investigating Alzheimer’s Disease using machine learning, the need to use both behavioural and biological markers data, and a computational method to rank Alzheimer’s Disease risk factors by importance using different machine learning models on Alzheimer’s Disease clinical assessment data from ADNI. The dataset contains Alzheimer’s Disease risk factors data related to medical history, family dementia history, demographical, and some lifestyle data for 1635 subjects. There are 387 normal control, 87 significant memory concerns, 289 early mild cognitive impairment, 539 late mild cognitive impairment and 333 Alzheimer’s Disease subjects. We deployed different machine learning models on the dataset to rank the importance of the variables (risk factors). The results show that some risk factors in subjects genetically, demography and lifestyle are more important than some medical history risk factors. Having APOE4, education level, age, weight, family dementia history, and type of work rank as more influential among Alzheimer’s Disease subjects.

In conclusion this study is aimed to explore the ADNI dataset and not to claim that the summary of it is medically of significance. The experiment in this study was conducted on a dataset which was extracted from a larger dataset provided by Alzheimer’s Disease Neuroimaging Initiative (ADNI). All subjects where unidentifiable, and the extracted data contained information that is related to Alzheimer’s Disease possible risk factors. The risk factors are categorised as behavioural risk factors such as lifestyle, demographic and characteristics, and biological risk factors such as medical history, genetics, and symptoms of sickness. The experiment employed four different machine learning models on the dataset extract to rank variables based on importance. These models include the Random Forest (RF), Neural Networks with a Principal Component Analysis (pcaNNet), Support Vector Machines with Linear Kernel (svmLinear), and Multi-Layer Perceptron (MLP). The overall output of the models combined suggests that all variables had some sort of importance but variables that had highest importance are energy, vision, dizziness, depression and headaches **[8].**

**2.9 PREDICTING DIABETES MELLITUS WITH MACHINE LEARNING TECHNIQUES**

**Q. Zou et al.,** has proposed in this work Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world’s diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. In order to verity the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments. We randomly selected 68994 healthy people and diabetic patients’ data, respectively as training set. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. In this study, we used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality.

The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used. Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using mRMR have better results. The result, which only used fasting glucose, has a better performance especially in Luzhou dataset. It means that the fasting glucose is the most important index for predict, but only using fasting glucose cannot achieve the best result, so if want to predict accurately, we need more indexes. In addition, by comparing the results of three classifications, we can find there is not much difference among random forest, decision tree and neural network, but random forests are obviously better than the another classifiers in some methods. The best result for Luzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721, which can indicate machine learning can be used for prediction diabetes, but finding suitable attributes, classifier and data mining method are very important. Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes **[9].**

**2.10 MOMENTUM PRINCIPAL SKEWNESS ANALYSIS**

**Xiurui Geng, Lingbo Meng et al.,** has proposed in this work Principal skewness analysis (PSA) has been introduced to the remote sensing community recently, which is equivalent to fast independent component analysis (FastICA) when skewness is considered as a non-Gaussian index. However, similar to FastICA, PSA also has the nonconvergence problem in searching for optimal projection directions. In this letter, we propose a new iteration strategy to alleviate PSA’s nonconvergence problem, and we name this new version of PSA as momentum PSA (MPSA). MPSA still adopts the same fixed-point algorithm as PSA does. Different from PSA, the (k + 1)th result in the iteration process of MPSA not only depends on the kth iteration result but also is related to the (k − 1)th iteration. Experiments conducted for both simulated data and real-world hyperspectral image demonstrate that MPSA has an obvious advantage over PSA in convergence performance and computational speed. In this letter, we have proposed an improved iteration method for PSA to search for the locally maximum skewness direction of an image, which is called MPSA. The only difference between MPSA and PSA is their iteration strategy. In PSA, the next iteration result is only dependent on the current iteration, which makes PSA easily trapped in oscillation situation.

However, the next iteration result of MPSA is determined by not only the current iteration result but also the last one. Both simulated and real-world experiments demonstrated that MPSA can greatly alleviate the divergence problem in PSA. In addition, MPSA also outperforms PSA in the computational efficiency. Similar to FastICA and PSA, MPSA can be applied to any multivariate data. However, when the statistical structure of targets of interest cannot be disclosed by skewness, MPSA will not be a good choice. It is noticeable that the common bottleneck of both PSA and MPSA is the calculation of coskewness tensor, which is quite time consuming. More effort should be placed on ways to compute the coskewness tensor more efficiently in the future. Additionally, both MPSA and PSA are based on thirdorder statistics, i.e., algorithms based on fourth- or higher order statistics may be worth studying **[10].**

**2.11 PREDICTING ACUTE KIDNEY INJURY IN CANCER PATIENTS USING HETEROGENEOUS AND IRREGULAR DATA**

**Namyong Park , Eunjeong Kang et al.,** has proposed in this work how can we predict the occurrence of acute kidney injury (AKI) in cancer patients based on machine learning with serum creatinine data? Given irregular and heterogeneous clinical data, how can we make the most of it for accurate AKI prediction? AKI is a common and significant complication in cancer patients, and correlates with substantial morbidity and mortality. Since no effective treatment for AKI still exists, it is important to take timely preventive measures. While several approaches have been proposed for predicting AKI, their scope and applicability are limited as they either assume regular data measured over a short hospital stay, or do not fully utilize heterogeneous data. In this work, we provide an AKI prediction model with a greater applicability, which relaxes the constraints of existing approaches, and fully utilizes irregular and heterogeneous data for learning the model. In a cohort of 21,022 cancer patients who were registered into Korea Central Cancer Registry (KCCR) in Seoul National University Hospital between January 1, 2004 and December 31, 2013, our method achieves 0.7892 precision, 0.7506 recall, and 0.7576 F-measure in predicting whether a patient will develop AKI during the next 14 days.

In this work, we present our method for predicting acute kidney disease (AKI) from irregular and heterogeneous clinical data. Our contributions are threefold. First, we present a model for AKI prediction that relaxes the constraints of existing prediction models, and thus has a better applicability. Second, we present a detailed explanation of our method. In particular, we delineate how we preprocess the heterogeneous and irregular data, how we create feature vectors, and how we structure our proposed framework for AKI prediction. Third, we provide extensive empirical evidences for the performance of our approach using the dataset of 21,022 cancer patients. Our method achieves 0.7892 precision, 0.7506 recall, and 0.7576 F-measure in predicting whether a patient will develop AKI during the next 14 days **[11].**

**2.12 USING RESISTIN, GLUCOSE, AGE AND BMI TO PREDICT THE PRESENCE OF BREAST CANCER**

**Miguel Patrício, José Pereira et al.,** has proposed in this work The goal of this exploratory study was to develop and assess a prediction model which can potentially be used as a biomarker of breast cancer, based on anthropometric data and parameters which can be gathered in routine blood analysis. For each of the 166 participants several clinical features were observed or measured, including age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP-1. Machine learning algorithms (logistic regression, random forests, support vector machines) were implemented taking in as predictors different numbers of variables. The resulting models were assessed with a Monte Carlo Cross-Validation approach to determine 95% confidence intervals for the sensitivity, specificity and AUC of the models Based on Resistin, Glucose, Age and BMI, the presence of breast cancer in women could be predicted on a test data set with sensitivity ranging between 82 and 88% and specificity ranging between 85 and 90% (95% CI for the AUC is [0.87, 0.91]). This suggests that Resistin and Glucose, taken together with Age and BMI, may be Fig. 5 ROC curves corresponding to the best and worst Support Vector Machine (SVM) models generated with four predictors in the cross-validation procedure Patrício et al. BMC Cancer (2018) 18:29 Page 7 of 8 considered a good set of candidates for breast cancer biomarkers to implement into screening tests.

As this procedure intends to increase the ease of diagnosis of breast cancer, it may potentially have great impact on the health of many women. In this study we propose a model for breast cancer detection based on biomarkers. The putative biomarkers assessed were Glucose, Resistin, Age, BMI, HOMA, Leptin, Insulin, Adiponectin, MCP-1. Using solely the combination of the first 4 variables on a predictive model using support vector machines allowed achieving the following 95% confidence intervals for sensitivity and specificity on a test set: [82%, 88%] and [84%, 90%], respectively. Additionally, the confidence interval for the AUC was [0.87, 0.91]. The intention is not to propose these models as an alternative to digital mammography, which a large study showed to have a sensitivity of 41% and a specificity of 98% at detecting which women would present breast cancer within 455 days of study entry and 70% sensitivity and 92% specificity when the follow up was reduced to 365 days [26]. Rather, as it is a rather noninvasive and inexpensive test which can be easily implemented in routine analysis by further measuring resistin (commercial kits allowing for the measurement of resistin are already available for under 20 euros per sample) and which we believe merits further study **[12].**

**2.13 A NEW EFFECTIVE MACHINE LEARNING FRAMEWORK FOR SEPSIS DIAGNOSIS**

**Xianchuan Wang, Zhiyi Wang et al.,** has proposed in this work there is a lack of early specific diagnosis and effective evaluation of sepsis, and the clinical treatment is not timely. As a result, the mortality is high, which seriously threatens the health of the people. Data were collected from the human blood samples of the hospital by gas chromatography mass spectrometry (GC-MS). 35 healthy controls and 42 sepsis patients were enrolled. Machine-learning techniques were used to diagnose the sepsis. Using the metabolic data from the sepsis patients, the proposed method has got 81.6% recognition rate, and 89.57% sensitivity and 65.77% specificity. A new learning strategy was proposed to boost the performance of kernel extreme learning machine (KELM), known as, chaotic fruit fly optimization (CFOA), and two new mechanisms were introduced into the original FOA, including the chaotic population initialization and chaotic local search strategy. To further enhance the diagnosis accuracy and identify the most important biomarkers, we performed the feature selection using the random forest (RF) before the construction of the classification model. The final established model, random forest-improved fruit fly optimization algorithm-kernel extreme learning machine (RF-CFOA-KELM), was used to effectively diagnose the sepsis. Experimental results demonstrate the proposed method obtains better results than other methods across four performance metrics.

We screened out five biomarkers and performed statistical analysis on these five substances. The level of acetic acid increased We proposed to use KELM to diagnose sepsis. Using the metabolic data from the sepsis patients, the proposed method has achieved a predictive accuracy of 81.6%. To enhance the performance of KELM, a new learning mechanism, the CFOA is proposed. We introduced two new mechanisms into the original FOA, including the chaotic population initialization and the chaotic local search strategy. To further promote the classification performance and identify the most important biomarkers, we performed the feature selection using the random forest prior to construct the classification model. The final established model, RF-CFOA-KELM, was used to diagnose the sepsis in an effective manner. We used the RF to screen out five biomarkers (D-xylose, acetic acid, linoleic acid, D-glucopyranosiduronic acid, and cholesterol), and we performed statistical analysis on these five substances **[13].**

**2.14 EARLY PREDICTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING SUPPORTED BY PREDICTIVE ANALYTICS**

**Ahmed J. Aljaaf1, 2, Dhiya Al-Jumeily et al.,** has proposed in this work Chronic Kidney Disease is a serious lifelong condition that induced by either kidney pathology or reduced kidney functions. Early prediction and proper treatments can possibly stop, or slow the progression of this chronic disease to end-stage, where dialysis or kidney transplantation is the only way to save patient’s life. In this study, we examine the ability of several machine-learning methods for early prediction of Chronic Kidney Disease. This matter has been studied widely; however, we are supporting our methodology by the use of predictive analytics, in which we examine the relationship in between data parameters as well as with the target class attribute. Predictive analytics enables us to introduce the optimal subset of parameters to feed machine learning to build a set of predictive models. This study starts with 24 parameters in addition to the class attribute, and ends up by 30% of them as ideal sub set to predict Chronic Kidney Disease. A total of 4 machine learning based classifiers have been evaluated within a supervised learning setting, achieving highest performance outcomes of AUC 0.995, sensitivity 0.9897, and specificity 1.

The experimental procedure concludes that advances in machine learning, with assist of predictive analytics, represent a promising setting by which to recognize intelligent solutions, which in turn prove the ability of predication in the kidney disease domain and beyond. To investigate ability of machine-learning, supported by predictive analysis, for early predication of CKD, an experimental procedure has undertaken in this study, considering a dataset collected from Apollo Hospitals—India, containing 400 instances. Two class labels used as targets in the study (i.e. patients with CKD and healthy individuals), over which four machine-learning methods were simulated. The classification and regression tree, i.e. RPART model, showing considerably good result. It uses the ratio of information gain for splitting criterion, where the optimal spilt would decrease impurity of resulting subsets. In this study, RPART stopping criterion of splitting was five, which means that next split will not occur **[14].**

**2.15 COMPARISON OF VARIABLE SELECTION METHODS FOR CLINICAL PREDICTIVE MODELING**

**L. Nelson Sanchez-Pintoa , Laura Ruth Venable et al.,** has proposed in this work Objective: Modern machine learning-based modeling methods are increasingly applied to clinical problems. One such application is in variable selection methods for predictive modeling. However, there is limited research comparing the performance of classic and modern for variable selection in clinical datasets. Materials and Methods: We analyzed the performance of eight different variable selection methods: four regression-based methods (stepwise backward selection using p-value and AIC, Least Absolute Shrinkage and Selection Operator, and Elastic Net) and four tree-based methods (Variable Selection Using Random Forest, Regularized Random Forests, Boruta, and Gradient Boosted Feature Selection). We used two clinical datasets of different sizes, a multicenter adult clinical deterioration cohort and a single center pediatric acute kidney injury cohort. Method evaluation included measures of parsimony, variable importance, and discrimination. Results: In the large, multicenter dataset, the modern tree-based Variable Selection Using Random Forest and the Gradient Boosted Feature Selection methods achieved the best parsimony. In the smaller, single-center dataset, the classic regression-based stepwise backward selection using p-value and AIC methods achieved the best parsimony.

In both datasets, variable selection tended to decrease the accuracy of the random forest models and increase the accuracy of logistic regression models. Conclusions: The performance of classic regression-based and modern tree-based variable selection methods is associated with the size of the clinical dataset used. Classic regression-based variable selection methods seem to achieve better parsimony in clinical prediction problems in smaller datasets while modern tree-based methods perform better in larger datasets. In conclusion, the performance of regression-based and tree-based variable selection methods is associated with the events-per-variable ratio of the clinical dataset used. Classic regression-based variable selection methods seem to achieve better parsimony in clinical prediction were normal, as performed in similar studies. Preliminary screening for collinearity was performed using pairwise correlations between all variables. The correlation for all pairs variables used in the analyses was < 0.75. A limited number of variable interactions were also explored preliminarily in a full logistic regression model. The correlation for all pairs variables used in the analyses was < 0.75. Interactions amongst variables were not explored in a logistic regression model given the constraints of the lower event-per-variable ratio **[15].**

**2.16 DIAGNOSIS OF CHRONIC KIDNEY DISEASE BY USING RANDOM FOREST**

**Abdulhamit Subas, Emina Alickovic et al.,** has proposed in this work Chronic kidney disease (CKD) is a global public health problem, affecting approximately 10% of the population worldwide. Yet, there is little direct evidence on how CKD can be diagnosed in a systematic and automatic manner. This work investigates how CKD can be diagnosed by using machine learning (ML) techniques. ML algorithms have been a driving force in detection of abnormalities in different physiological data, and are, with a great success, employed in different classification tasks. In the present study, a number of different ML classifiers are experimentally validated to a real data set, taken from the UCI Machine Learning Repository, and our findings are compared with the findings reported in the recent literature. The results are quantitatively and qualitatively discussed and our findings reveal that the random forest (RF) classifier achieves the near-optimal performances on the identification of CKD subjects. Hence, we show that ML algorithms serve important function in diagnosis of CKD, with satisfactory robustness, and our findings suggest that RF can also be utilized for the diagnosis of similar diseases. In this study, different machine learning methods are used for diagnosis of CKD. In this study, we employed different machine learning methods for diagnosis of chronic kidney disease (CKD).

Results represented with Confusion Matrix are shown in Table 1. Precision, F-measure, and Overall Classification Accuracy results are shown in Table 2. As it can be seen easily from these tables, the highest performances were obtained by using random forest (RF) classifier. Overall accuracy of RF is 100%. C4.5 decision tree classifier also resulted in high accuracy rate of 99 %. Support vector machine with setting parameter c equal to 100 and employing normalized poly kernel has achieved 98.5 % where 396 patterns were correctly classified. ANN classifier resulted overall accuracy rate of 98 %. K-NN classifier resulted in poorest overall accuracy rate of 95.75 % where 383 patterns were correctly classified. This study showed that RF classifier results in very high performances during classification tasks. This method is able to classify two different classes with perfect classification rate, 100%. Precision and Fmeasure for RF are also perfect, 1.0 and 1.0. We also applied ANN, k-NN, SVM and C4.5 decision tree methods and obtained high performance results, especially for C4.5 decision tree classifier. For diagnosis of CKD, we propose the usage of RF machine learning tool since it results in high classification accuracy rate, time needed for training and testing is low **[16].**

**2.17 DIAGNOSIS OF CHRONIC KIDNEY DISEASE BASED ON SUPPORT VECTOR MACHINE BY FEATURE SELECTION METHODS**

**Huseyin Polat 1 & Homay Danaei Mehr et al.,** has proposed in this work As Chronic Kidney Disease progresses slowly, early detection and effective treatment are the only cure to reduce the mortality rate. Machine learning techniques are gaining significance in medical diagnosis because of their classification ability with high accuracy rates. The accuracy of classification algorithms depend on the use of correct feature selection algorithms to reduce the dimension of datasets. In this study, Support Vector Machine classification algorithm was used to diagnose Chronic Kidney Disease. To diagnose the Chronic Kidney Disease, two essential types of feature selection methods namely, wrapper and filter approaches were chosen to reduce the dimension of Chronic Kidney Disease dataset. In wrapper approach, classifier subset evaluator with greedy stepwise search engine and wrapper subset evaluator with the Best First search engine were used. In filter approach, correlation feature selection subset evaluator with greedy stepwise search engine and filtered subset evaluator with the Best First search engine were used. The results showed that the Support Vector Machine classifier by using filtered subset evaluator with the Best First search engine feature selection method has higher accuracy rate (98.5%) in the diagnosis of Chronic Kidney Disease compared to other selected methods. In this study, wrapper and filter methods have been utilized on data set of CKD.

Two different evaluators have been used for each method. For filter approach, CfsSubsetEval with Greedy stepwise search engine and FilterSubsetEval with Best First search engine have been used. In addition to wrapper approach, ClassifierSubsetEval with Greedy stepwise search engine and WrapperSubsetEval with Best First search engine have been used. The accuracy rate of SVM classifier on full training set has been compared with its accuracy rate on 4 reduced datasets which have been gained by feature selection methods. The results show that after reducing dimension of CKD dataset, in all 4 methods accuracy rate of diagnosis have been improved. The accuracy rate of SVM classification on reduced dataset by FilterSubsetEval with Best First search engine (98.5%) is more than the other used methods. It has obtained the highest values of other comparable results such as TP rate, correctly classified instances. Moreover, it has lowest number of important values such as incorrectly classified instances and FP rate. It is noticeable that, preparing the dataset with the lowest dimension by feature selection methods could not lead to the highest accuracy rate of classification in perpetuity. For instance, the accuracy rate of SVM on the lowest dimension of CKD dataset by 7 attributes by ClassifierSubsetEval with Greedy stepwise search engine is not the highest accuracy rate (98%), however the accuracy rate of SVM classifier on 13 attributes of CKD dataset, by using FilterSubsetEval with Best First feature selection method, has got the most accuracy rate (98.5%) in CKD diagnosis. Furthermore, with different methods of feature selection and classification algorithms, on distinct datasets of disease, classification results can be different in accuracy rates **[17].**

**2.18 CHRONIC KIDNEY DISEASE, BASAL INSULIN GLARGINE AND HEALTH OUTCOMES IN PEOPLE WITH DYSGLYCEMIA: THE ORIGIN STUDY**

**Vasilios Papademetriou, Michael Doumas et al.,** has proposed in this work Early stages of chronic kidney disease are associated with an increased cardiovascular risk in patients with established type 2 diabetes and macrovascular disease. The role of early stages of chronic kidney disease on macrovascular outcomes in prediabetes and early type two diabetes mellitus is not known. In the Outcome Reduction with an Initial Glargine Intervention (ORIGIN) trial, the introduction of insulin had no effect on cardiovascular outcomes compared to standard therapy. In this post hoc analysis of ORIGIN, we compared cardiovascular outcomes in subjects without to those with mild (Stages 1-2) and/or moderate chronic kidney disease. Τwo co-primary composite cardiovascular outcomes were assessed. The first was the composite endpoint of non-fatal MI, non-fatal stroke, or death from cardiovascular causes; and the second was a composite of any of these events plus a revascularization procedure, or hospitalization for heart failure. Several secondary outcomes were pre-specified including microvascular outcomes, incident diabetes, hypoglycemia, weight, and cancers. Complete renal function data were available in 12,174 out of 12,537 ORIGIN participants. A total of 8,114 had no chronic kidney disease (67%) while 4,060 had chronic kidney disease stage 1-3 (33%). When compared to non-CKD participants, the risk of developing the composite primary outcome (nonfatal myocardial infarction, nonfatal stroke, or cardiovascular death) in those with mild to moderate chronic kidney disease was 87% higher; hazard ratio (HR): 1.87; 95% confidence interval (CI): 1.71-2.04 (p<0.001).

This subgroup analysis of the ORIGIN data clearly identifies both mild (chronic kidney disease 1 -2) and moderate chronic kidney disease (stage 3) as significant risk factors for macrovascular outcomes in people with early dysglycemia. Moreover, these results were not impacted by the use of basal insulin glargine compared to standard treatment. The adverse role of mild to moderate chronic kidney disease on macrovascular outcomes are very similar to a previous analysis of the ACCORD trial, where subjects of similar age but with more advanced diabetes mellitus and higher baseline HgA1c were recruited.8 In both trials, however, the rate of macrovascular complication at baseline, was high and of a similar magnitude. These findings on chronic kidney disease and outcome concordance are reinforced by the ADVANCE study with a similar degree of baseline macrovascular prevalence.9 Although both the baseline glycemic status and glycemic management goals in these trials (i.e., ACCORD and ADVANCE vs ORIGIN) were significantly different, the role of the mild to moderate chronic kidney disease vis-à-vis the macrovascular outcomes did not differ. Interestingly, in a diabetes cohort of 1,493 patients with diabetic chronic kidney disease, with a lower incidence of macrovascular outcomes than either ACCORD, ORIGIN or ADVANCE, the authors found a significant increase in macrovascular disease prevalence from stage 3 (17.78%) to 4 (52.48%) and a critical cut off level of eGFR at 46.4 ml/min/1.73m2 . 14 Thus, the adverse impact of early chronic kidney disease appears to be related to the baseline macrovascular status, irrespective of the degree of dysglycemia **[18].**

**2.19 MACHINE-LEARNING-BASED CLASSIFICATION OF REAL-TIME TISSUE ELASTOGRAPHY FOR HEPATIC FIBROSIS IN PATIENTS WITH CHRONIC HEPATITIS**

**Yang Chen, Yan Luo, Wei Huang et al.,** has proposed in this work Hepatic fibrosis is a common middle stage of the pathological processes of chronic liver diseases. Clinical intervention during the early stages of hepatic fibrosis can slow the development of liver cirrhosis and reduce the risk of developing liver cancer. Performing a liver biopsy, the gold standard for viral liver disease management, has drawbacks such as invasiveness and a relatively high sampling error rate. Real-time tissue elastography (RTE), one of the most recently developed technologies, might be promising imaging technology because it is both noninvasive and provides accurate assessments of hepatic fibrosis. However, determining the stage of liver fibrosis from RTE images in a clinic is a challenging task. In this study, in contrast to the previous liver fibrosis index (LFI) method, which predicts the stage of diagnosis using RTE images and multiple regression analysis, we employed four classical classifiers (i.e., Support Vector Machine, Naïve Bayes, Random Forest and K-Nearest Neighbor) to build a decision-support system to improve the hepatitis B stage diagnosis performance. Eleven RTE image features were obtained from 513 subjects who underwent liver biopsies in this multicenter collaborative research.

The experimental results showed that the adopted classifiers significantly outperformed the LFI method and that the Random Forest(RF) classifier provided the highest average accuracy among the four machine algorithms. This result suggests that sophisticated machine-learning methods can be powerful tools for evaluating the stage of hepatic fibrosis and show promise for clinical applications. In conclusion, this study applied four classical pattern-recognition methods, Naïve Bayes, RF, KNN and SVM, to build a decision-support system to estimate the stage of hepatic fibrosis based on real-time ultrasonic tissue elastography. The results reported here indicate that machine learning could be a superior method for staging hepatitis fibrosis compared to statistical methods. In addition, we believe that the proposed models could be helpful tools to assist specialists in determining the stage of hepatic fibrosis. The algorithm can be adapted for ultrasound diagnostic instruments such as the Hitachi-HI VISION Preirus (HITACHI Medical Corporation, Tokyo Japan), and the pathological classification can be estimated to provide doctors with faster and more reliable diagnostic approaches for these diseases **[19].**

**2.20 PERFORMANCE EVALUATION ON MACHINE LEARNING CLASSIFICATION TECHNIQUES FOR DISEASE CLASSIFICATION AND FORECASTING THROUGH DATA ANALYTICS FOR CHRONIC KIDNEY DISEASE (CKD)**

**Gunarathne, Perera et al.,** has proposed in this work Chronic Kidney Disease (CKD) is considered as kidney damage which lasts longer than three months. In Sri Lanka, CKD has become a severe problem in the present days due to CKD of unknown aetiology (CKDu) that can be seen popularly in North Central Province. Identifying CKD in the initial stage is important to provide necessary treatments to prevent or cure the disease. In this work main focus is on predicting the patient’s status of CKD or non CKD. To predict the value in machine learning classification algorithms have been used. Classification models have been built with different classification algorithms will predict the CKD and non CKD status of the patient. These models have applied on recently collected CKD dataset downloaded from the UCI repository with 400 data records and 25 attributes. Results of different models are compared. From the comparison it has been observed that the model with Multiclass Decision forest algorithm performed best with an accuracy of 99.1% for the reduced dataset with the 14 attributes. As an advantage of that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a data mining model targeting chronic kidney disease with overall accuracy = 99.99%, will need thousands and thousands of records with zero missing values. But at this level of the study, it will demonstrate the strength of the models by observing statistics metric values which are acceptable in data mining. When considering the decision tree algorithm it builds the tree based on the entire dataset by using all the features of the dataset but, the decision forest builds using multiple decision trees by categorizing the features of the dataset in to separate trees. The tree with the highest probability have the greater effect on the final result in decision forest. Because of that, decision forest gives more accurate prediction result than the decision tree. Rather than using the decision tree, by using the above model created using the decision forest algorithm it can achieve a 99.1% of better accuracy with less number of attributes **[20].**

MATERIALS AND METHODS

**3.1 EXISTING SYSTEM**

The data identified with the undertaking what's more, acquires the qualities of the relating design. This innovation can accomplish exact and practical analyses of sicknesses; subsequently, it very well may be a promising technique for diagnosing CKD.

The existing system predicts the chronic diseases which are for a particular region and for the particular community. Only particular diseases are predicted by this system. In this System, Big Data & CNN Algorithm is used for Disease risk prediction. For S type data, the system is using Machine Learning algorithm i.e Decision Tree, Naïve Bayesian. The accuracy of the existing System is up to 94.8%.

In the existing work, they streamline machine learning algorithms for the effective prediction of chronic disease outbreak in disease-frequent communities. They experiment with the modified prediction models over reallife hospital data collected from central China. They propose a convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from the hospital. It has gotten another sort of clinical instrument with the improvement of data innovation what's more, has an expansive application prospect in view of the fast improvement of electronic wellbeing record . In the clinical field, has just been utilized to detect human body status break down the significant components of the infection and analyze different sicknesses. For instance, the models worked by machine learning calculations were utilized to analyze coronary illness , diabetes and retinopathy , intense kidney injury , disease what's more, different sicknesses. In these models, calculations in view of relapse, tree, likelihood, choice surface and neural organization were regularly compelling.

**3.1.1 DISADVANTAGES:**

* This is mainly due to each statement in the test contributing a single data point in measuring a facet of the underlying trait.
* Therefore, the disease may not be detected until the kidney loses about 25% of its function.
* As a result, their method could not be used when the diagnostic results of the samples are unknown.
* In reality, patients might miss some measurements for various reasons before diagnosing.
* In addition, for missing values in categorical variables, data obtained using mean imputation might have a large deviation from the actual values.

## 3.2 PROPOSED SYSTEM

They used picture enrollment to recognize renal morphologic changes and set up a classifier dependent on neural organization utilizing enormous scope CKD information, and the exactness of the model on their test information. Moreover, the majority of the past examines used the CKD informational index that was acquired from the UCI AI store. They utilized k-closest neighbor (KNN), uphold vector machine and delicate autonomous displaying of class relationship to analyze KNN and Logistic regression system for diagnosing CKD accomplished the most elevated exactness of 100% Chronic kidney disease (CKD) is a global public health problem, affecting approximately 10% of the population worldwide. Yet, there is little direct evidence on how CKD can be diagnosed in a systematic and automatic manner.

This work investigates how CKD can be diagnosed by using machine learning (ML) techniques. ML algorithms have been a driving force in detection of abnormalities in different physiological data, and are, with a great success, employed in different classification tasks. In the present study, a number of different ML classifiers are experimentally validated to a real data set, taken from the UCI Machine Learning Repository, and our findings are compared with the findings reported in the recent literature. The results are quantitatively and qualitatively discussed and our findings reveal that the Logistic regression (LR) classifier achieves the near-optimal performances on the identification of CKD subjects. Hence, we show that ML algorithms serve important function in diagnosis of CKD, with satisfactory robustness, and our findings suggest that LR can also be utilized for the diagnosis of similar diseases.Their examinations have accomplished great outcomes in the finding of CKD. In the above models, the mean ascription is utilized to fill in the missing qualities and it relies upon the demonstrative classifications of the examples. Therefore, their technique couldn't be utilized at the point when the demonstrative consequences of the examples are obscure. In reality, patients may miss a few estimations for different reasons prior to diagnosing.

**3.2.1 ADVANTAGES:**

* It is used to validate the accuracy of the trained models.
* After misjudgement analysis, the better algorithms that produce different misjudgements are extracted as component models.
* An integrated model is then established to improve the performance of the classifier.
* We believe that this model will be more and more perfect by the increase of size and quality of the data.
* Missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy..

**3.3 MODULE DESCRIPTION**

* **DATA PROCESSING**
* **EXTRACTING FEATURE SELECTION**
* **PERFORMANCE INDICATORS**
* **ESTABLISHING AND EVALUATING INDIVIDUAL MODELS**
* **MISJUDGMENT ANALYSIS AND SELECTING COMPONENT MODELS**

**DATA PROCESSING**

Each categorical (nominal) variable was coded to facilitate the processing in a computer. For the values of rbc and pc, normal and abnormal were coded as 1 and 0, respectively. For the values of pcc and ba, present and not present were coded as 1 and 0, respectively. For the values of htn, dm, cad, pe and ane, yes and no were coded as 1 and 0, respectively. For the value of appet, good and poor were coded as 1 and 0, respectively. Although the original data description denes three variables sg, al and su as categorical types, the values of these three variables are still numeric based, thus these variables were treated as numeric variables. All the categorical variables were transformed into factors. Each sample was given an independent number that ranged from 1 to 400. There is a large number of missing values in the data set, and the number of complete instances is 158. In general, the patients might miss some measurements for various reasons before making a diagnosis. Thus, missing values will appear in the data when the diagnostic categories of samples are unknown, and a corresponding imputation method is needed.

**EXTRACTING FEATURE SELECTION**

Extracting feature vectors or predictors could remove variables that are neither useful for prediction nor related to response variables and thus prevent these unrelated variables the models to make an accurate prediction . Here in, we used optimal subset regression and LR to extract the variables that are most meaningful to the prediction. Optimal subset regression detects the model performance of all possible combinations of predictors and selects the best combination of variables. LR detects the contribution of each variable to the reduction in the Gini index. The larger the Gini index, the higher the uncertainty in classifying the samples. Therefore, the variables with contribution of 0 are treated as redundant variables. The step of feature extraction was run on each complete data set The combinations are ranked from left to right by the degree The vertical axis represents variables. The horizontal axis is the adjusted r-squared which represents the degree to which the combination of variables explains the response variable. To make it easy to distinguish each combination of variables, we used four colors (red, green, blue and black) to mark the selected variables. The combinations are ranked from left to right by the degree of explanations to the response variable and the right-most combination has the strongest interception to the response variable.

**K-NN CLASSIFICATION:**

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In K-NN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbor.

In a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance.

The limitation of the K-NN algorithm is it’s sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction.

Chronic Kidney Disease dataset is taken from UCI database which consists of 25 variables with 400 instances. In that we have continuous, nominal and binary variables. Hence nominal variables attributes such as specific gravity, albumin and sugar are taken. We convert all the nominal variables to binary and we use knn classification. k values are chosen.

In the training phase, a KNN algorithm is applied and in the test phase results are displayed.

The main goal of this work is to evaluate the performance of ten distance formulae when KNN is used for binary data and also to find the best value of k. Here, we assign for k values ranging from 175 to 190 and find out the resulting error rates.

**LOGISTIC REGRESSION:**

Logistic regression: Logistic regression also called logic model or logistic model, is a widely used model to analyze the relationship between multiple independent variables and one categorical dependent variable

The logistic regression model can work well for regression, but fails for classification. Why is that? In case of two classes, you could label one of the classes with 0 and the other with 1 and use linear regression. Technically it works and most linear model programs will spit out weights for you. But there are a few problems with this approach:

A logistic model does not output probabilities, but it treats the classes as numbers (0 and 1) and fits the best hyperplane (for a single feature, it is a line) that minimizes the distances between the points and the hyperplane. So it simply interpolates between the points, and you cannot interpret it as probabilities. A linear model also extrapolates and gives you values below zero and above one. This is a good sign that there might be a smarter approach to classification. Since the predicted outcome is not a probability, but a linear interpolation between points, there is no meaningful threshold at which you can distinguish one class from the other.

**PERFORMANCE INDICATORS**

In this study, ckd was set to be positive and notckd was set to be negative. The confusion matrix was used to show the specific results and evaluate the performance of the machine learning models. True positive (TP) indicates the ckd samples were correctly diagnosed. False negative (FN) indicates the ckd samples were incorrectly diagnosed. False positive (FP) indicates the notckd samples were incorrectly diagnosed. True negative (TN) indicates the notckd samples were correctly diagnosed. Accuracy, sensitivity, specicity, precision, recall and F1 score were used to evaluate the performance of the model. They are calculated using the following equations.

Accuracy =

Recall = sensitivity =

specificity =

precision =

F1 score =

**ESTABLISHING AND EVALUATING INDIVIDUAL MODELS**

The following machine learning models have been obtained by using the corresponding subset of features or predictors on

the complete CKD data sets for diagnosing CKD.

1) Logistic Regression-based model:LR

2) Distance-based model: KNN

Generally, in disease diagnosis, diagnostic samples are distributed in a multidimensional space. This space comprises predictors that are used for data classi\_cation (ckd or notckd). Samples of data in the space are clustered in different regions due to their different categories. Therefore, there is a boundary between the two categories, and the distances between samples in the same category are smaller. According to the effectiveness of classification, we choose the aforementioned methods for disease diagnosis. LR is based on Logistic Regression, and it obtains the weight of each predictor and a bias. If the sum of the effects of all predictors exceeds a threshold, the category of the sample will be classified as ckd or notckd. RF generates a large number of decision trees by randomly sampling training samples and predictors. Each decision tree is trained to find a boundary that maximises the difference between ckd and notckd. The final decision is determined by the predictions of all trees in the disease diagnosis. Divides different kinds of samples by establishing a decision surface in a multidimensional space that comprises the predictors of the samples. KNN finds the nearest training samples by calculating the distances between the test sample and the training samples and then determines the diagnostic category by voting. Naive Bayes classifier calculates the conditional probabilities of the sample under the interval by the number of ckd and notckd samples in each different measurement interval. KNN can analyse non-linear relationships

in the data sets due to its complex structure, and the sigmoid activation function was used in the hidden layer and the output layer.

**MISJUDGMENT ANALYSIS AND SELECTING COMPONENT MODELS**

After evaluating the above models, the potential component models were extracted for misjudgement analysis to determine which would be used as the components. The misjudgement analysis here refers to nd out and compare the samples misjudged by different models, and then determine which model is suitable to establish the nal integrated model. The misjudgement analysis was performed on the extracted models. The prerequisite for generating an integrated model is that the misjudged samples from each component model are different. If each component model misjudges the same samples, the generated integrated model would not make a correct judgement for the samples either. When the data were read, each sample was given a unique number ranging from 1 to 400.

The numbers of misjudgements for the extracted models on each complete data and the black part indicates that the samples were misjudged by other models except KNN. LR), when K equalling to 7, only one misjudgement is simultaneously misjudged by the LR. In other cases, all the samples that are misjudged by LR can be correctly judged by the rest of the models. Hence, the combinations of the LR with the rest of the models could be used to establish an integrated model. Next, we investigate which specific model combination could generate the best integrated model for diagnosing CKD.

**3.4 SYSTEM REQUIREMENTS**

**3.4.1 HARDWARE REQUIREMENTS:**

* Processor Type ` : Pentium i3
* Speed : 3.40GHZ
* RAM : 4GB DD2 RAM
* Hard disk : 500 GB
* Keyboard : 101/102 Standard Keys
* Mouse : Optical Mouse

**3.4.2 SOFTWARE REQUIREMENTS:**

* Operating System : Windows 10
* Front end : Jupyter Notebook / Pycharm IDE
* Coding Language : Python
* Tools : Anaconda Navigator

**CHAPTER 4**

EXPERIMENTAL SETUP AND PROCEDURE

To evaluate model performance comprehensively, in the case of retaining the sample distribution in the original data, a complete data set was divided into four subsets evenly. For all of the above models, each subset was utilized once for testing, and other subsets were utilized for training, the overall result was taken as the final performance.

In order to verify whether the integrated model can improve the performance of the component models, we first used the same random number seed 1234 to establish and evaluate the integrated model on each complete data, and the confusion matrices returned. Comparing, it can be found that the integrated model improves the performance of the component models and achieves an accuracy of 100% when K equaling to 3 and 11.When K equaling to 5, 7 and 9, the integrated model improves the performance of LR and has the same accuracy. Next, for a comprehensive evaluation, we removed the random number seed 1234 which was used to divide the data into four subsets and establish the LR.

The integrated model was then run 10 times on the complete data sets. The average results of the integrated models and two component models. Our results show the feasibility of the proposed methodology. By the use of KNN imputation, LR, could achieve better performance than the imputation was used. KNN imputation could fill in the missing values in the data set for the cases wherein the diagnostic categories are unknown, which is closer to the real-life medical situation. Through the misjudgments analysis, LR were selected as the component models. The LR achieved an accuracy of around 98.75%, which indicates most samples in the data set are linearly separable. The KNN achieved better performance compared with the LR with the accuracy was around 99.75%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Class** | **Precision** | **F-**  **Measure** | **Accuracy (%)** |
| **LR** | **CKD** | 0.996 | 0.988 | 98 |
| **No-CKD** | 0.968 | 0.980 | 99.3 |
| **Weighted Average** | 0.985 | 0.985 | 98.5 |
| **K-NN** | **CKD** | 1.000 | 0.965 | 93.2 |
| **No-CKD** | 0.898 | 0.946 | 100 |
| **Weighted Average** | 0.962 | 0.958 | 95.75 |

**TABLE 1.1** Integrated model for Acc ,Sen, Spec and F1 represent the accuracy,sensitivity,specifity and F1 score, respectively.

**GRAPH 4.1** INTEGRATED MODEL FOR ACC ,SEN, SPEC AND F1

# CHAPTER 5

RESULTS AND DISCUSSION

We speculate that this methodology could be extended to more complex situations. When processing more complex data, various different algorithms are attempted to establish models. After misjudgement analysis, the better algorithms that produce different misjudgements are extracted as component models. An integrated model is then established to improve the performance of the classifier. It can be seen that the proposed methodology improves the performance of the otherwise independent models and achieves comparable or better performance compared to the models proposed in previous studies.

In addition, the CKD data set is composed of mixed variables (numeric and category), so the similarity evaluation methods based on mixed data could be used to calculate the similarity between samples, such as general similarity coefficient. In this study, we used euclidean distance to evaluate the similarity between samples, and KNN could obtain a good result based on Euclidean distance with the highest accuracy of 100%. Therefore, we did not use on the methods to evaluate the similarity between samples.

CHAPTER 6

CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. In this assessment, we propose an KNN and Logistic regression, system for diagnosing CKD Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of establishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples.

Therefore, the generalization performance of the model might be limited. In addition, due to there are only two categories (ckd and notckd) of data samples in the data set, the model cannot diagnose the severity of CKD.In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data.

**REFERENCES**

1. M. M. Hossain et al., “Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts,” IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Mar. 2019.
2. E. Hodneland et al., “In vivo detection of chronic kidney disease using tissue deformation fields from dynamic MR imaging,” IEEE Trans. BioMed. Eng., vol. 66, no. 6, pp. 1779-1790, Jun. 2019.
3. G. R. Vasquez-Morales et al., “Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning,” IEEE Access, vol. 7, pp. 152900-152910, Oct. 2019.
4. N. Almansour et al., “Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study,” Comput. Biol. Med., vol. 109, pp. 101-111, Jun. 2019
5. M. Alloghani et al., “Applications of machine learning techniques for software engineering learning and early prediction of students’ performance,” in Proc. Int. Conf. Soft Computing in Data Science, Dec. 2018, pp. 246- 258.
6. L. Du et al., “A machine learning based approach to identify protected health information in Chinese clinical text,” Int. J. Med. Inform., vol. 116, pp. 24-32, Aug. 2018
7. R. Abbas et al., “Classification of foetal distress and hypoxia using machine learning approaches,” in Proc. Int. Conf. Intelligent Computing, Jul. 2018, pp. 767-776
8. M. Mahyoub, M. Randles, T. Baker and P. Yang, “Comparison analysis of machine learning algorithms to rank alzheimer’s disease risk factors by importance,” in Proc. 11th Int. Conf. Developments in eSystems Engineering, Sep. 2018.
9. Q. Zou et al., “Predicting diabetes mellitus with machine learning techniques,” Front. Genet., vol. 9, Nov. 2018
10. Z. Gao et al., “Diagnosis of diabetic retinopathy using deep neural networks,” IEEE Access, vol. 7, pp. 3360-3370, Dec. 2018.
11. N. Park et al., “Predicting acute kidney injury in cancer patients using heterogeneous and irregular data,” Plos One, vol. 13, no. 7, Jul. 2018
12. M. Patricio et al., “Using resistin, glucose, age and BMI to predict the presence of breast cancer,” BMC CANCER, vol. 18, Jan. 2018
13. X. Wang et al., “A new effective machine learning framework for sepsis diagnosis,” IEEE Access, vol. 6, pp. 48300-48310, Aug. 2018
14. A. J. Aljaaf et al., “Early prediction of chronic kidney disease using machine learning supported by predictive analytics,” in Proc. IEEE Congr. Evolutionary Computation, Jul. 2018.
15. L. N. Sanchez-Pinto, L. R. Venable, J. Fahrenbach, M. M. Churpek, “Comparison of variable selection methods for clinical predictive modeling,” Int. J. Med. Inform., vol. 116, pp. 10-17, Aug. 2018
16. A. Subasi, E. Alickovic, J. Kevric, “Diagnosis of chronic kidney disease by using random forest,” in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.
17. H. Polat, H.D. Mehr, A. Cetin, “Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,” J. Med. Syst., vol. 41, no. 4, Apr. 2017
18. V. Papademetriou et al., “Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study,” Am. J. Med., vol. 130, no. 12, Dec. 2017
19. Y. Chen et al., “Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B,” Comput. Biol. Med., vol. 89, pp. 18-23, Oct. 2017
20. W. H. S. D. Gunarathne, K. D. M. Perera and K. A. D. C. P. Kahandawaarachchi, “Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD),” in Proc. IEEE 17th Int. Conf. Bioinformatics and Bioengineering, Oct. 2017, pp. 291-296