(Project Proposal for MSc)

Date: April 14, 2022

1. **Name of the student:** N. I. MD. ASHAFUDDULA
   Student ID: 18204016
2. **Tentative Title:** A COMPREHENSIVE ANALYSIS TO PREDICT CHRONIC KIDNEY DISEASE EFFICIENTLY AT AN EARLY STAGE USING MACHINE LEARNING ALGORITHMS
3. **Background and present state of the problem:** Chronic kidney disease (CKD) is defined as the progressive and irreversible damage to the kidneys that, over the course of months or years, can lead to kidney (renal) failure [1]. CKD is one of the most critical illnesses nowadays and proper diagnosis is required as soon as possible. As there is no cure for CKD, there are treatments that can significantly slow the progression of the disease if started early [1]. Studies (9 studies, a total of 225,206 participants) based on meta-analysis showed an overall prevalence of CKD in Bangladeshi people of 22.48%, which was higher than the global prevalence of CKD [2]. The prevalence of CKD in females was higher with high heterogeneity (I2 90%) in contrast to male participants (25.32% vs. 20.31%) [2]. Machine Learning (ML) technique has become reliable for medical treatment. With the help of Machine Learning classifier algorithms the doctors can detect the disease on time. From this perspective, CKD has been chosen.

   Authors (2021) [3] applied Data pre-processing techniques such as Data encoding and Missing values filled up. The missing values are handled by the researchers with four different criteria like mean, mode, median and null dropping method. RandomizedSearchCV is used to automate hyper-parameter tuning. Eight ML algorithms are used to predict CKD where Random Forest 99.75% produced better accuracy than any other classifier. In this work, authors have not shown result analysis on the feature selection, dimensionality reduction and handling of imbalance data in the dataset. Moreover, their work have not been evaluated on clinical data. Authors (2020) [4] omitted missing values from the dataset, they have used feature selection techniques and 11 ML algorithms to predict CKD. In this work, authors split the dataset into 70% train set, 15% validation and 15%test set where 70% train set could overfit the model. They achieved 100% accuracy using Decision Tree, Random Forest, Extra Trees Classifier and ADA Boost Classifier. Authors excluded a few features manually without showing proper effectiveness of them on the model and also clinical data analysis has not shown. Dataset pre-processing is done before feeding them into the ML and Deep learning (DL) model by the authors (2021) [5].They also used Feature selection methodologies and solved class imbalance problem in the dataset by using Synthetic Minority Oversampling Technique (SMOTE). The authors achieved best result using LSVM with penalty L2: 98.86% and DNN: 99.6%. They also did not show model performance on clinical data.
4. **Objectives with specific aims and possible outcome:**

   The objectives of this work are enumerated below:

   i. To develop a model to effectively analysis the CKD.
   ii. Efficiently use of data encoding techniques such as Dummy Encoding.
   iii. Feature selection methods such as L1 based feature selection and Variance Threshold based feature selection will be used.
   iv. Principle Component Analysis (PCA) will be used to reduce feature space.
   v. To Solve the data imbalance problem by using sampling technique.
   vi. Logistic Regression, Naive Bayes, Decision Tree, Random Forest, KNN, SVM, Passive Aggressive classifier, Gradient Boost, Stochastic Gradient Descent supervised ML classifiers will be used.
   vii. The model performance will be evaluated on clinical the dataset.

   The possible outcomes of the proposed research work are as follows:

   i. This work will help to understand to analysis the performance of effectively selection of features.
   ii. It can be used to detect whether the patient has CKD or not.
   iii. The overall performance of the model will be improved compared to the existing model.
   iv. Clinical data analysis will be shown.

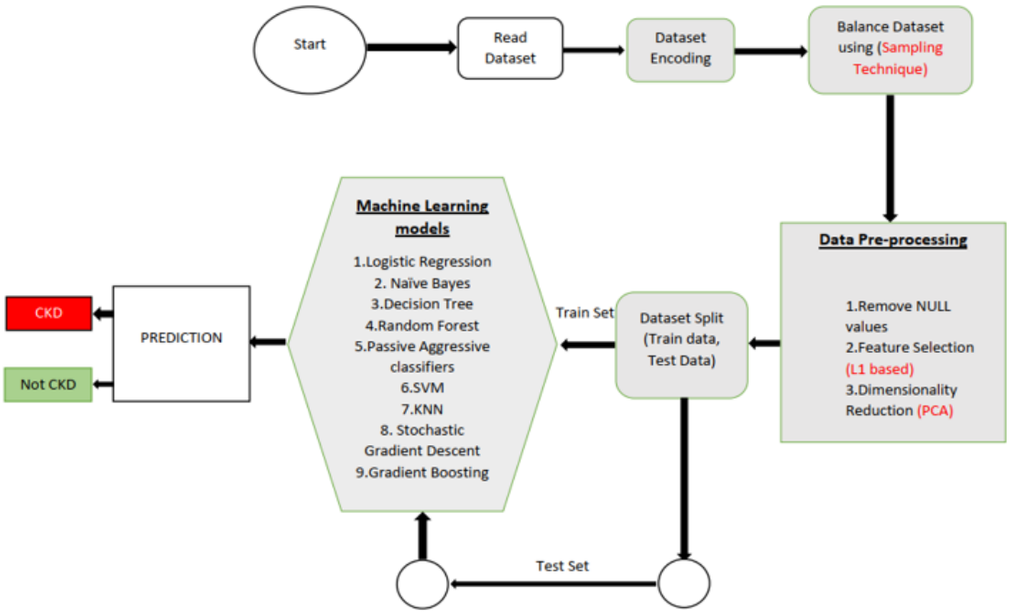

Figure 1: Proposed methodology

5. **Proposed Methodology** Fig. 1 shows the overall proposed methodology of this work.

6. **Challenges:** In this work, the main challenge is to work with raw data. The CKD dataset contains a lot of missing values, categorical variables and text which need to be well pre-processed before feeding into the model. To get better performance here we need to focus on the pre-processing of the dataset thus, the proposed solution outperforms the existing ML model performance.

7. **References:**

[1] J. Myhre and D. Sifris, "How chronic kidney disease is treated," 2021. https://www.verywellhealth.com/kidney-disease-treatments-4170060.

[2] S. Banik and A. Ghosh, "Prevalence of chronic kidney disease in bangladesh: A systematic review and meta-analysis," *International Urology and Nephrology*, vol. 53, no. 4, pp. 713–718, 2021.

[3] M. M. Nishat, R. R. Dip, F. Faisal, S. M. Nasrullah, R. Ahsan, M. F. Shikder, M. A.-A.-R. Asif, and M. A. Hoque, "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 18, no. e6, 2021.

[4] I. U. Ekanayake and D. Herath, "Chronic kidney disease prediction using machine learning methods," in *2020 Moratuwa Engineering Research Conference (MERCon)*, pp. 260–265, IEEE, 2020.

[5] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasiński, Ł. Jasiński, R. Gono, E. Jasińska, *et al.*, "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021.