A COMPREHENSIVE ANALYSIS TO PREDICT CHRONIC KIDNEY DISEASE EFFICIENTLY AT AN EARLY STAGE USING MACHINE LEARNING ALGORITHMS

Presented by---

Name: N. I. Md. Ashafuddula

Student ID: 18204016

Department: CSE

Program: MSc in CSE

Outlines

- ❑Introduction
- ❑Literature Review
- ❑Motivation
- ❑Proposed Methodology
- ❑Result Discussion
- ❑Future work
- ❑Conclusion
- ❑References

Introduction

- 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].
- There is no cure for CKD, there are treatments that can significantly slow the progression of the disease if started early [1].

- The treatment can vary based on your stage of disease and the underlying cause, such as Diabetes or High blood pressure.

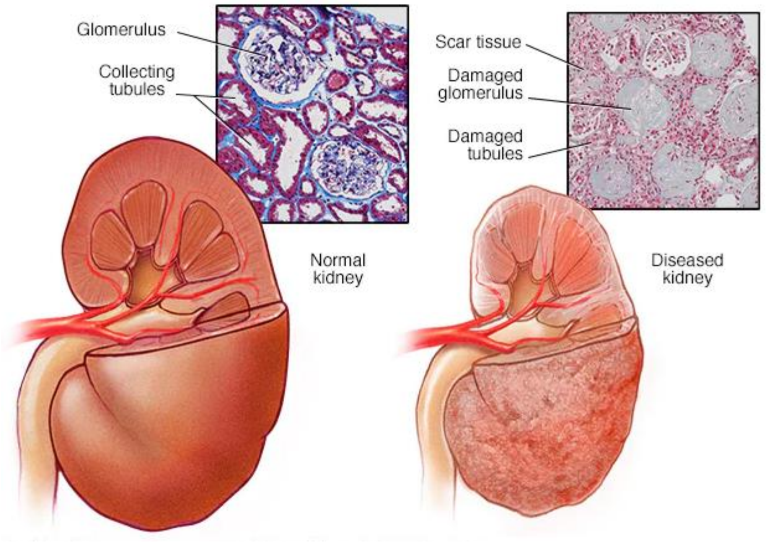

Fig 1.: Healthy kidney vs. diseased kidney [2]

Introduction (Cont'd)

- 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 [3].
- The prevalence of CKD in females was higher with high heterogeneity (I2 90%) in contrast to male participants (25.32% vs. 20.31%) [3].

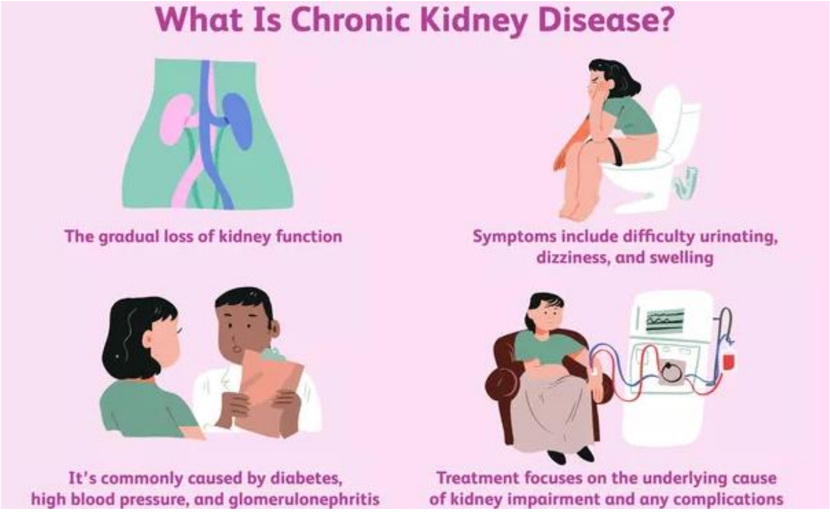

Fig 2.: Chronic Kidney Disease[1]

Literature Review

Reference Papers	**Proposed Methodology**	**Highest result** (accuracy)
A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms (2021) [4]	1. Data preprocessing **(Data encoding, Missing values filled up)** *2.* ***RandomizedSearchCV*** is used to automate hyperparameter tuning 3. Used 8 Machine Learning algorithms	Random Forest: 99.75%
Prediction of chronic kidney disease-a machine learning perspective (2021)[5]	1. Dataset preprocessing **2. Feature selection** 3. Classifier application **4. Solved class imbalance problem** in the dataset by using Synthetic Minority Oversampling Technique (SMOTE) 5. Analyzing the performance of the classifier	LSVM with penalty L2: 98.86% **DNN**: 99.6%
Chronic Kidney Disease Prediction Using Machine Learning Methods (2020) [6]	**1. Missing value omitted** 2. Used **feature Selection** techniques 3. Used 11 Machine Learning algorithms	1. Decision Tree, 2. Random Forest, 3. Extra Trees Classifier, 4. ADA Boost Classifier: 100%

Motivation

As these procedures,

- Did not mention any dimensionality reduction method so there is a high possibility of getting miss classification result due to overfitting the model as well as it causes an extra amount of times.

- Did not show performance of their model for Clinical new data

We need a computerized artificial intelligence-based system which can automatically detect and classify Chronic Kidney Disease at an early-stage with less amount of time and greater accuracy.

Proposed Methodology

- We will use a series of pre-processing steps in the dataset to reduce artifacts that could mislead the Machine Learning algorithms.

Proposed Methodology (Cont'd)

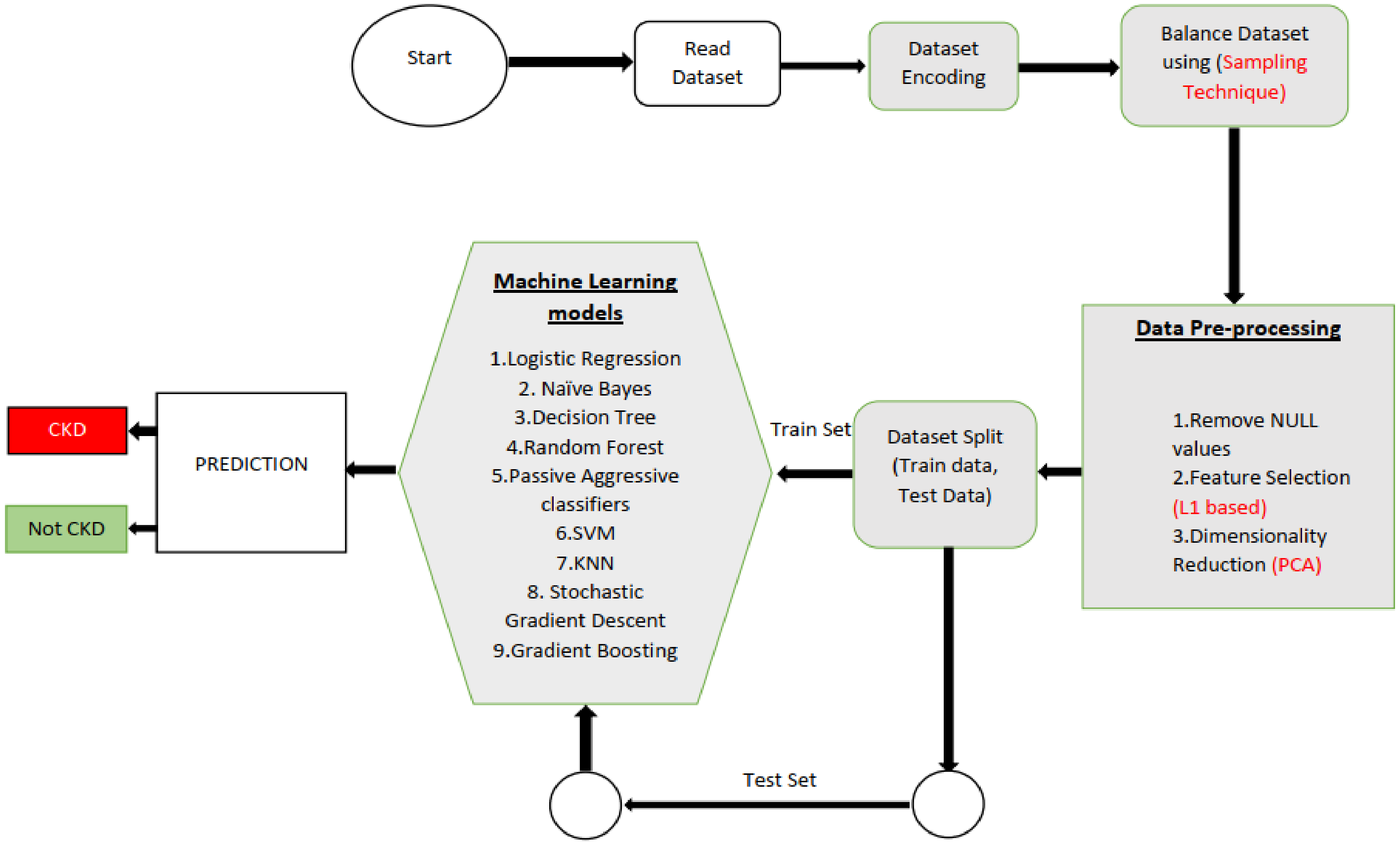

Fig 3.: Proposed methodology

Proposed Methodology (Cont'd)

- **Dataset**

  To evaluate this proposed methodology Dataset 2015[7] and 2021[8] are used.

- **Feature Selection Technique**

  Linear Support Vector Classification (LSVC) (with L1 penalty)

- **Dimension Reduction Technique**

  Principal component analysis (PCA)

Result Discussion

- Dataset 2015, **Data (503*25)**
- Selected features **13 out of 25**
- Dimension Reduction (PCA**) 2 from 13 features**

Table 1. Dataset-2015 Result Discussion

SL	Classifier name	Training Accuracy	Testing Accuracy	ROC-AUC
1	Logistic Regression	100	100	1.00
2	Decision Tree	100	100	1.00
3	Random Forest	100	100	1.00
4	Passive Aggressive Classifier	100	100	1.00
5	SVM	100	100	1.00
6	KNN	100	100	1.00
7	Gradient Boosting	100	100	1.00
8	Naïve Bayes	97.16	96.03	0.991
9	Stochastic Gradient Descent	94.6	94.04	0.941

Result Discussion (Cont'd)

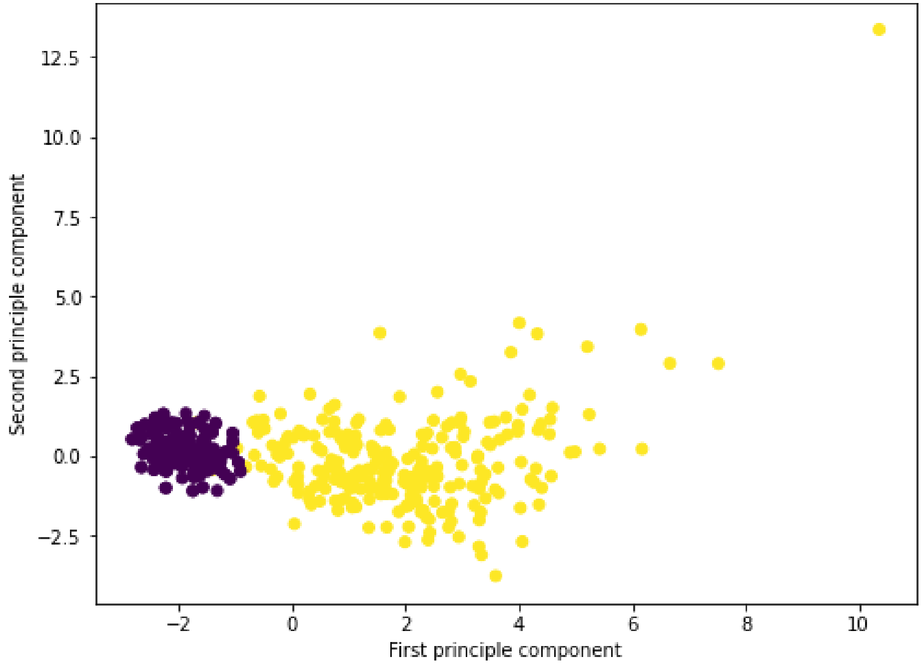

Fig 4. Dataset-2015 feature space (PCA = 2)

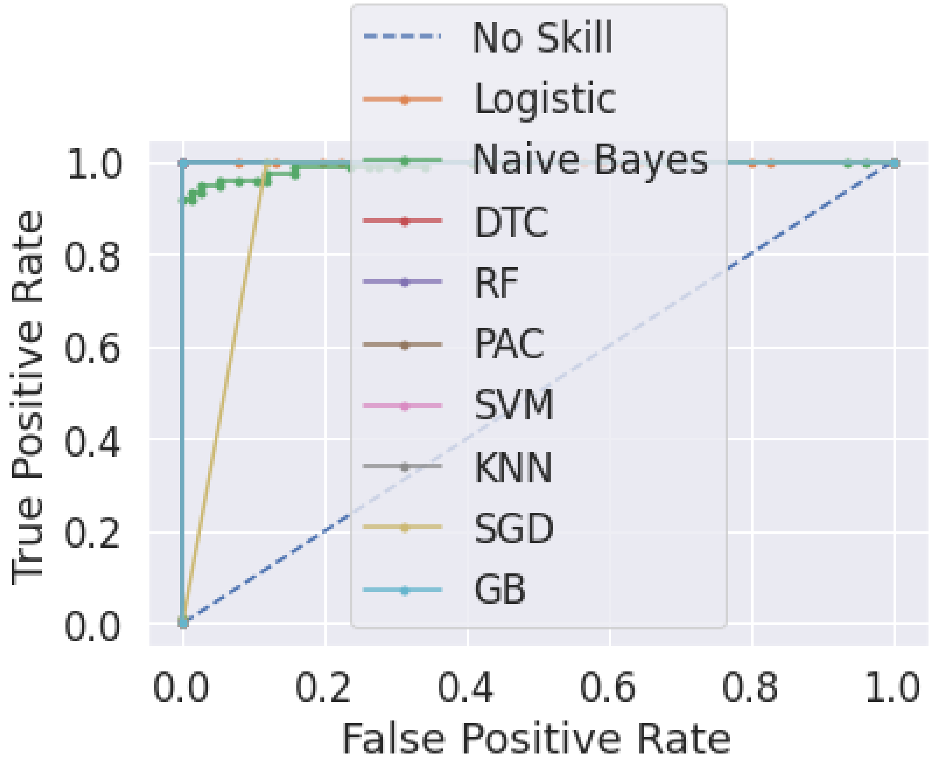

Fig 5. Dataset-2015 ROC-AUC curve

Result Discussion (Cont'd)

- Dataset 2021, **Data (256*27)**
- Used Categorical encoding, Remove **NaN** with Average value
- Selected features **16 out of 27**
- Dimension Reduction (PCA) **7 from 16 features**

Table 2. Dataset-2021 Result Discussion

SL	**Classifier name**	**Training Accuracy**	**Testing Accuracy**	**ROC-AUC**
1	Decision Tree	100	100	1.00
2	Random Forest	100	100	1.00
3	KNN	100	100	1.00
4	Gradient Boosting	100	98.70	0.987
5	Stochastic Gradient Descent	97.21	98.70	1.00
6	Naïve Bayes	98.70	98.70	0.981
7	SVM	98.32	97.40	0.974
8	Logistic Regression	97.21	96.1	0.997
9	Passive Aggressive Classifier	96.09	97.4	0.974

Result Discussion (Cont'd)

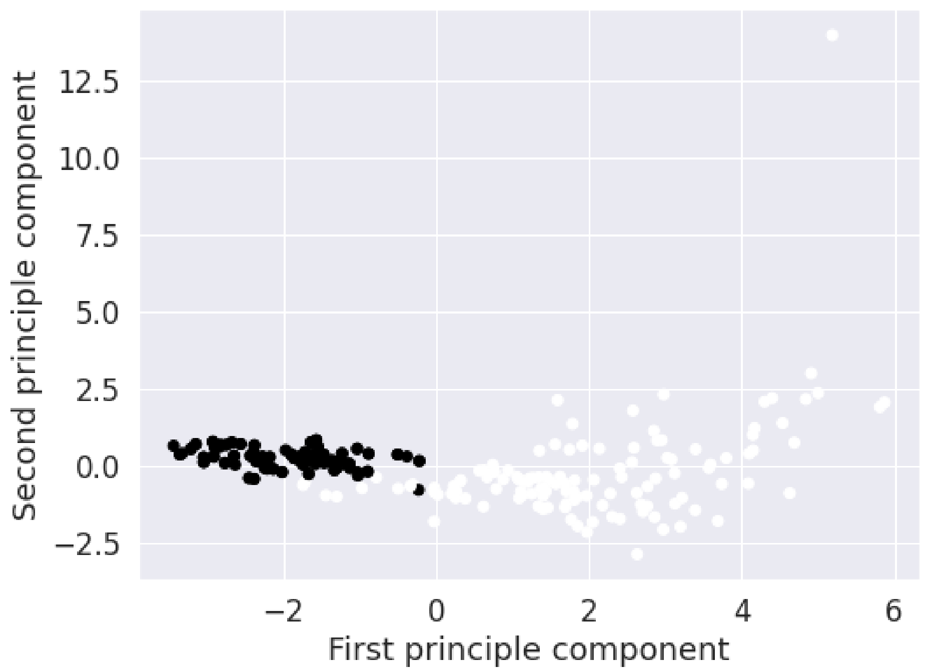

Fig 6. Dataset-2021 feature space (PCA = 7)

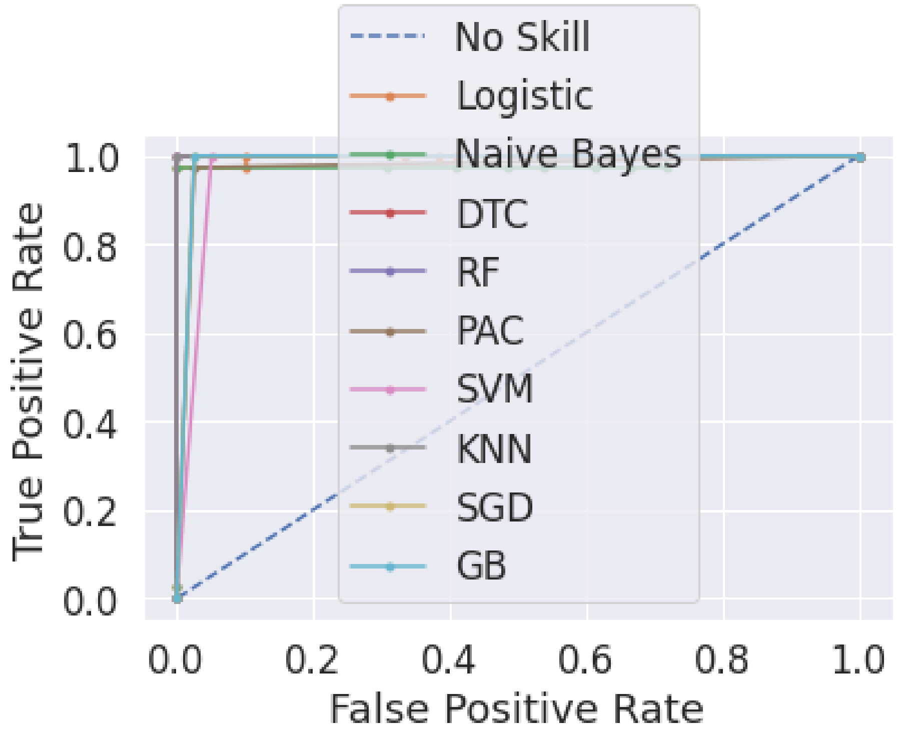

Fig 7. Dataset-2021 ROC-AUC curve

Result Discussion (Cont'd)

- Dataset 2015 (503*25) & 2021 (256*27)
- Selected features dataset 2015 **(503*13)** & 2021 **(256*14)**
- Dimension Reduction **(PCA = 3)**
- **Merge** two dataset **(759*3)**

Table 3. Hybrid Dataset Result Discussion

SL	Classifier name	Training Accuracy	Testing Accuracy	ROC-AUC
1	Gradient Boosting	98.87	98.25	0.982
2	Decision Tree	100	97.37	0.974
3	Random Forest	98.87	97.37	0.974
4	Passive Aggressive Classifier	97.93	97.37	0.974
5	SVM	98.31	97.37	0.974
6	Logistic Regression	97.55	96.93	0.994
7	KNN	97.55	96.05	0.961
8	Stochastic Gradient Descent	97.36	96.49	0.965
9	Naïve Bayes	96.80	95.61	0.986

Result Discussion (Cont’d)

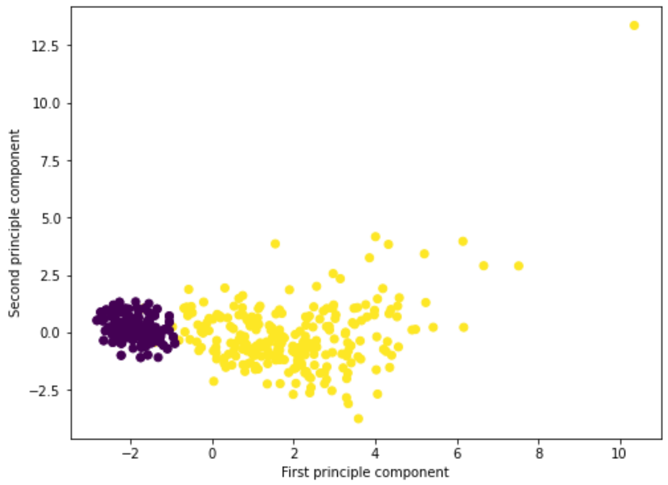

Fig 8. Dataset-15 feature space (PCA = 3)

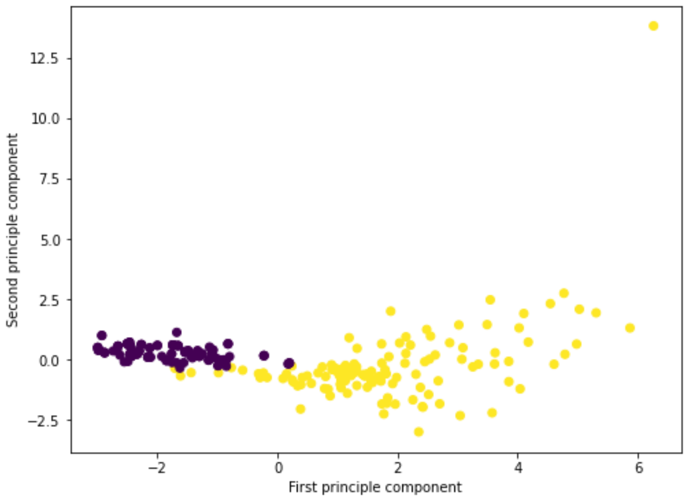

Fig 9. Dataset-21 feature space (PCA = 3)

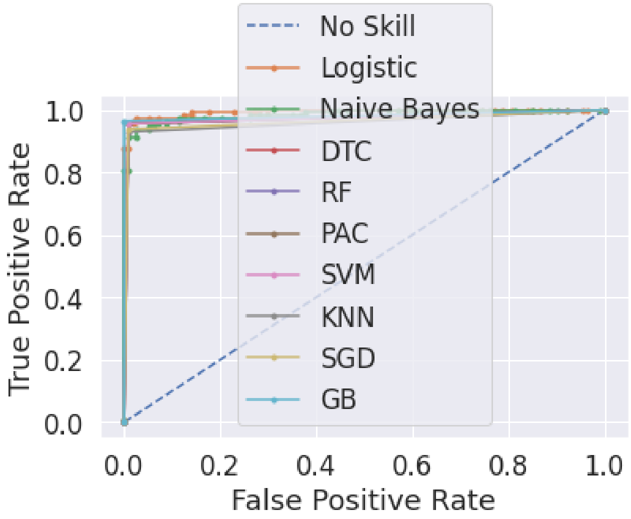

Fig 10. Dataset-2021 ROC-AUC curve

Result Discussion (Cont'd)

- Dataset 2015 (503*25) & 2021 (256*27)
- Selected features dataset 2015 **(503*13)** & 2021 **(256*14)**
- Dimension Reduction **(PCA = 10)**

Table 4. Clinical Unseen Result Discussion

SL	Classifier name	Training Accuracy	Testing Accuracy	ROC-AUC
1	Naïve Bayes*	97.22	95.7	0.980
2	SVM	99.2	95.31	0.953
3	Logistic Regression	99.01	94.92	0.985
4	KNN	98.41	94.53	0.945
5	Passive Aggressive Classifier	99.4	91.41	0.914
6	Random Forest	100	90.62	0.906
7	Decision Tree	100	88.42	0.824
8	Gradient Boosting	100	87.11	0.871
9	Stochastic Gradient Descent	99.4	84.77	0.848

Result Discussion (Cont'd)

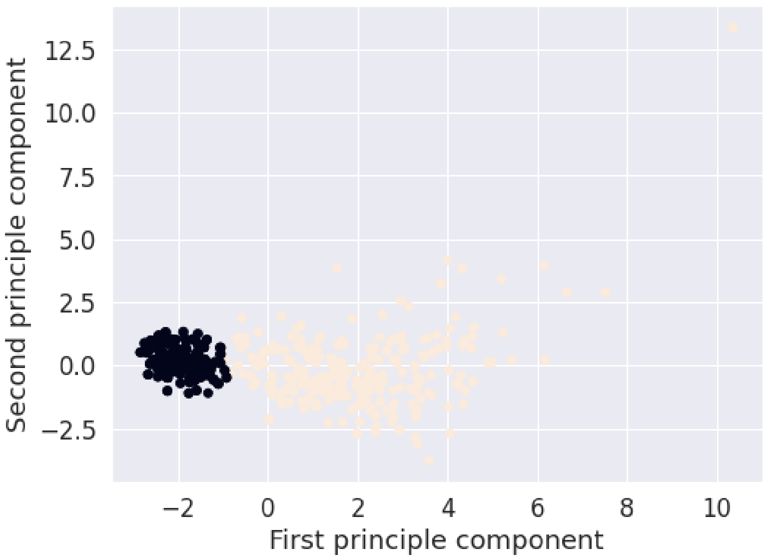

Fig 11. Dataset-15 feature space (PCA = 10)

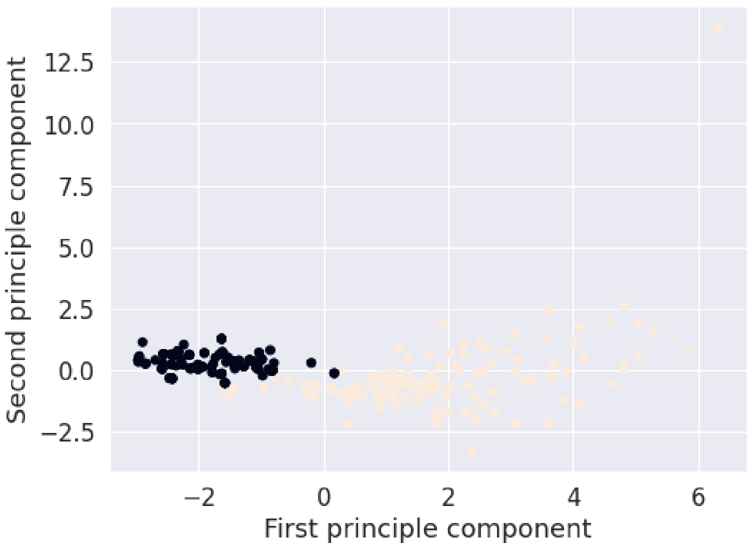

Fig 12. Dataset-21 feature space (PCA = 10)

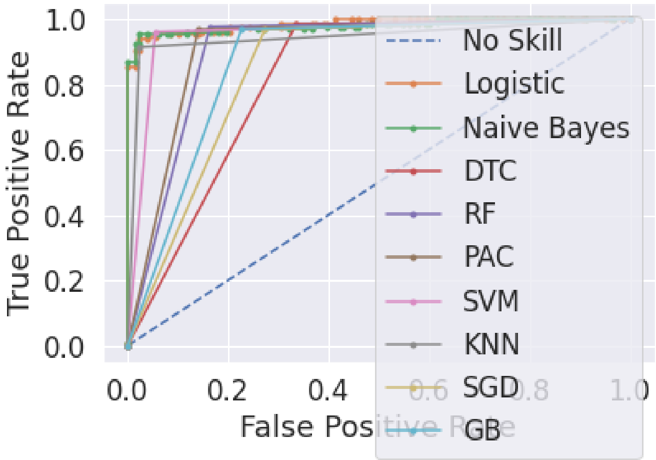

Fig 13. Unseen data ROC-AUC curve

Result Analysis

- Classifier that performs best for different data

Table 5. Result analysis for different data

Dataset-2015	Dataset-2021	Hybrid Dataset	Clinical Unseen Data
1. Logistic Regression **(100%)** 2. Decision Tree **(100%)** 3. Random Forest **(100%)** 4. Passive Aggressive Classifier **(100%)** 5. SVM **(100%)** 6. KNN **(100%)** 7. Gradient Boosting **(100%)** 8. Naïve Bayes **(95.7%)**	1. Decision Tree **(100%)** 2. Random Forest **(100%)** 3. KNN **(100%)** 4. Naïve Bayes **(98.70)** 5. SVM **(97.40)** 6. Logistic Regression **(96.1%)**	1. Gradient Boosting **(98.25%)** 2. Decision Tree **(97.37%)** 3. Random Forest **(97.37%)** 4. Passive Aggressive Classifier **(97.37%)** 5. SVM **(97.37%)** 6. Logistic Regression **(96.93%)** 7. KNN **(96.05%)** 8. Naïve Bayes **(95.61%)**	1. Naïve Bayes **(95.7%)** 2. SVM **(95.31%)** 3. Logistic Regression **(94.92%)** 4. KNN **(94.53%)** 5. Random Forest (90.62%) 6. Decision Tree (88.42%)

Result Analysis (Cont'd)

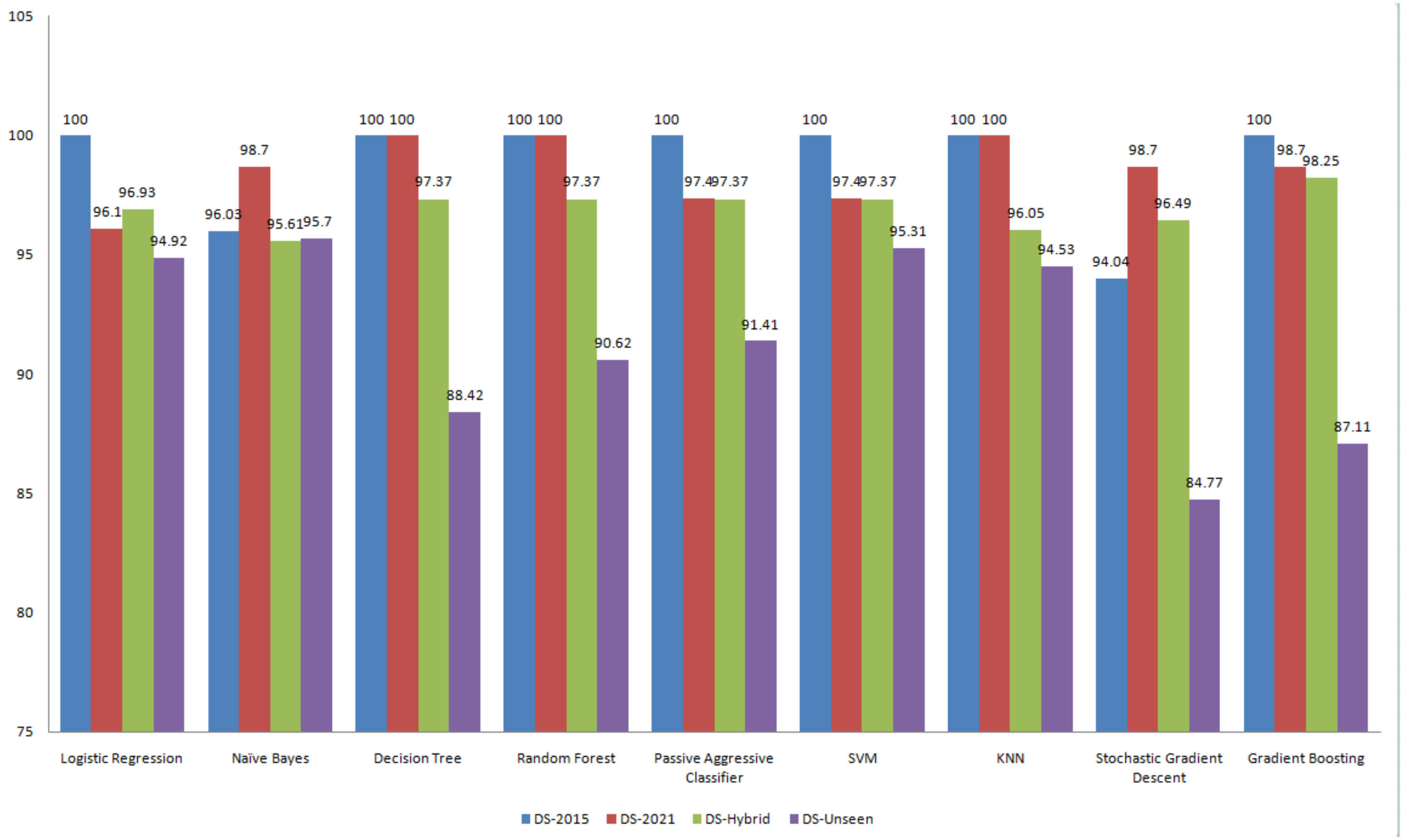

Fig 14. Result comparison

Result Analysis (Cont’d)

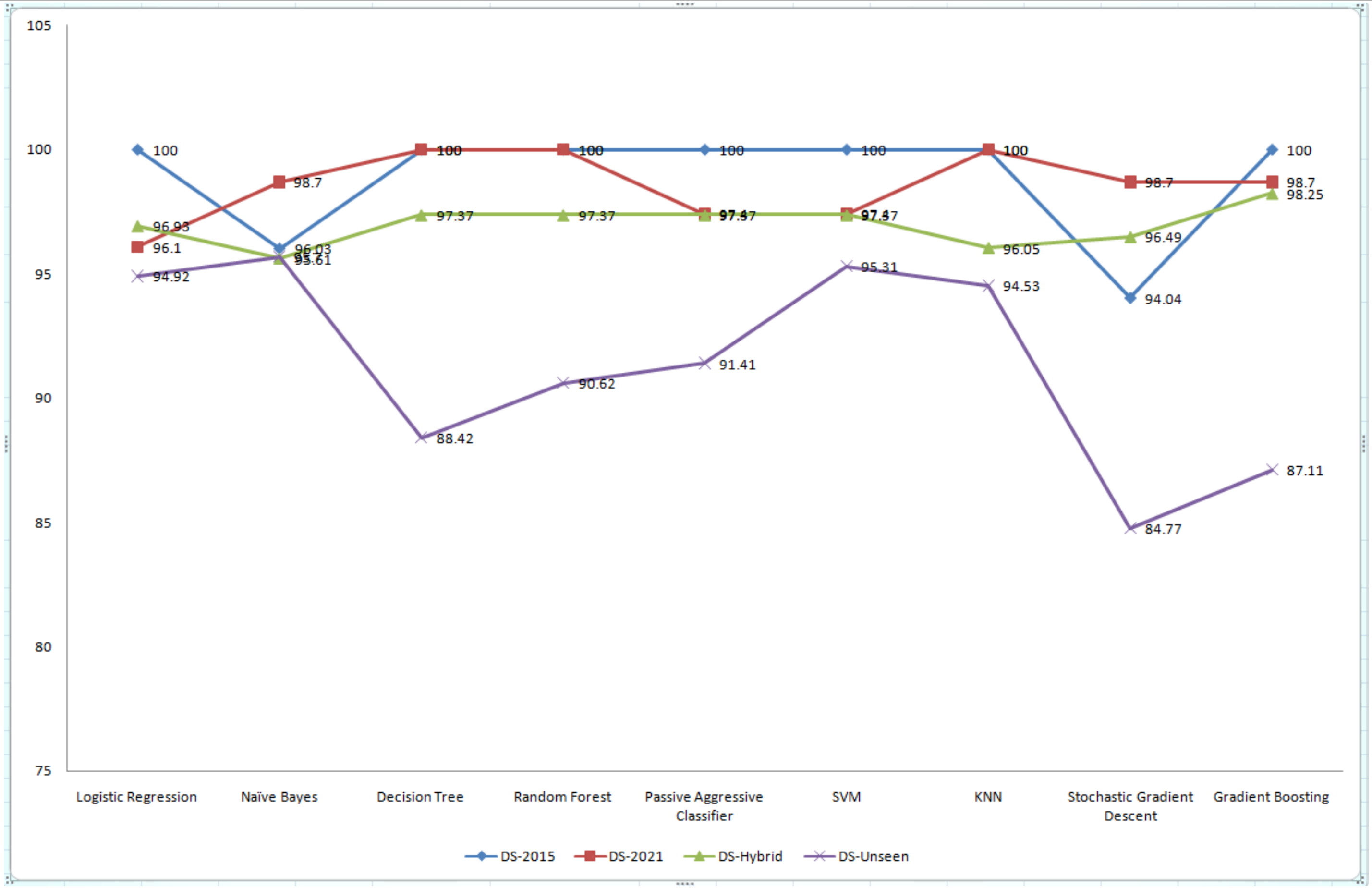

Fig 15. Result comparison

Future work

- Using other feature selection methods could be the possible future work.

Conclusion

- In this work, the main challenge is to work with raw data. The dataset contains a lot of missing values, categorical variables and text which need to be pre-processed before feeding into the model.
- To get better performance here we are focusing on the preprocessing of the dataset thus the proposed solution outperforms the existing Machine Learning model performance.

References

[1] J. Myhre, D. Sifris, How Chronic Kidney Disease Is Treated From Diet and Drugs to Dialysis and Transplant, very well health, October 11, 2021, Accessed on: April. 9, 2022. [Online]. Available: https://www.verywellhealth.com/kidney-disease-treatments-4170060

[2] Mayo C. Staff, Chronic kidney disease care at Mayo Clinic, Chronic kidney disease, Sept. 03, 2021, Accessed on: April. 6, 2022
[Online]. Available: https://libraryguides.vu.edu.au/ieeereferencing/webbaseddocument

[3] Banik, Sujan, and Antara Ghosh. "Prevalence of chronic kidney disease in Bangladesh: A systematic review and meta-analysis." *International Urology and Nephrology* 53, no. 4 (2021): 713-718.

[4] Nishat, Mirza Muntasir, Rezuanur Rahman Dip, Fahim Faisal, Sarker Md Nasrullah, Ragib Ahsan, Md Fahim Shikder, Md Asfi-Ar-Raihan Asif, and Md Ashraful Hoque. "A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms." *EAI Endorsed Transactions on Pervasive Health and Technology* 18, no. e6 (2021).

References

[5] Chittora, Pankaj, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Michał Jasiński et al. "Prediction of chronic kidney disease-a machine learning perspective." *IEEE Access* 9 (2021): 17312-17334.

[6] Ekanayake, Imesh Udara, and Damayanthi Herath. "Chronic kidney disease prediction using machine learning methods." In *2020 Moratuwa Engineering Research Conference (MERCon)*, pp. 260-265. IEEE, 2020.

[7] Rubini,L.Jerlin,UCIMachineLearningRepository[http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease]. Karaikudi,TamilNadu: Algappa University, Department of Computer Science and Engineering, 2015

[8] Islam, Md Ashiqul, Shamima Akter, Md Sagar Hossen, Sadia Ahmed Keya, Sadia Afrin Tisha, and Shahed Hossain. "Risk factor prediction of chronic kidney disease based on machine learning algorithms." In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 952-957. IEEE, 2020.

Thank You