Lung Cancer Risk Prediction Report

Humera Nuzhat
2303A52083
Batch 39

Introduction

The report gives an account of how and what a machine learning project based on predicting Chronic Kidney Disease (CKD) was carried out using a given clinical dataset. The main aim was two-fold, the first to create and test different machine and deep learning models in order to have a high predictive accuracy, and the second to use Explainable AI (XAI) methods to get knowledge about how the most successful models make decisions. This interpretability plays an important role in clinical uses, as it enables healthcare providers to believe and confirm the predictions of the model.

The dataset, kidney_disease.csv, is a combination of both numerical and categorical variables based on lab outcomes and clinical findings of patients. The target variable is classification, which implies the presence or absence of CKD.

Key Findings

Robust Models: Each of the five models was strong predictors on the cleaned data set.

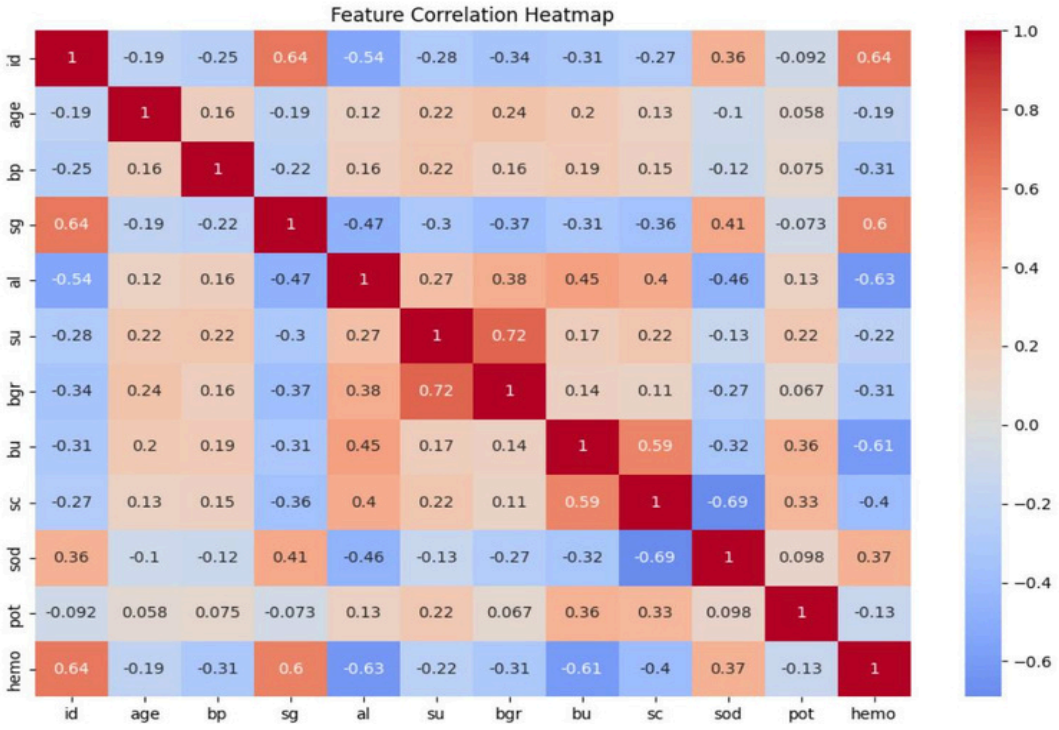
Near-Perfect Performance: The ensemble techniques (Random Forest and XGBoost) and the ANN had high scores that were close to 100 in terms of their accuracy.

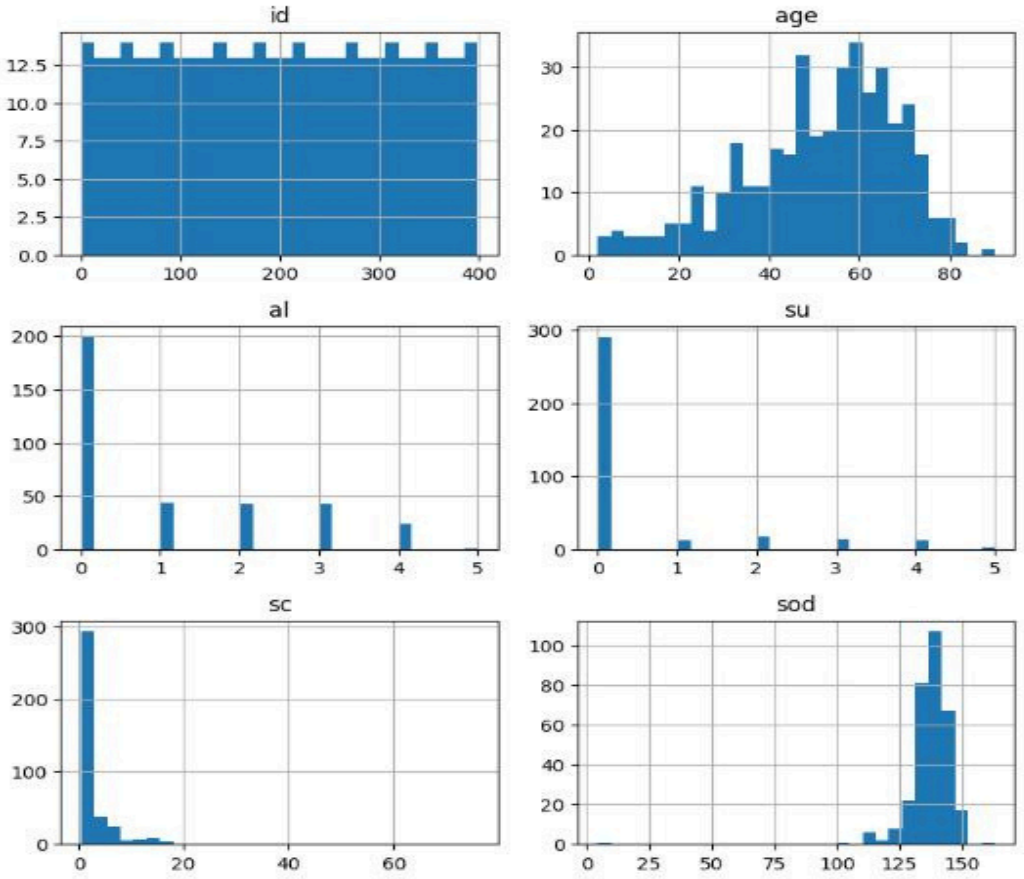
Critical Features: The XAI analysis showed that some key features were always the most contributing to the prediction of CKD in all the models. These were sg (specific gravity), al (albumin), rbc (red blood cells), sc (serum creatinine), and hemo (hemoglobin).

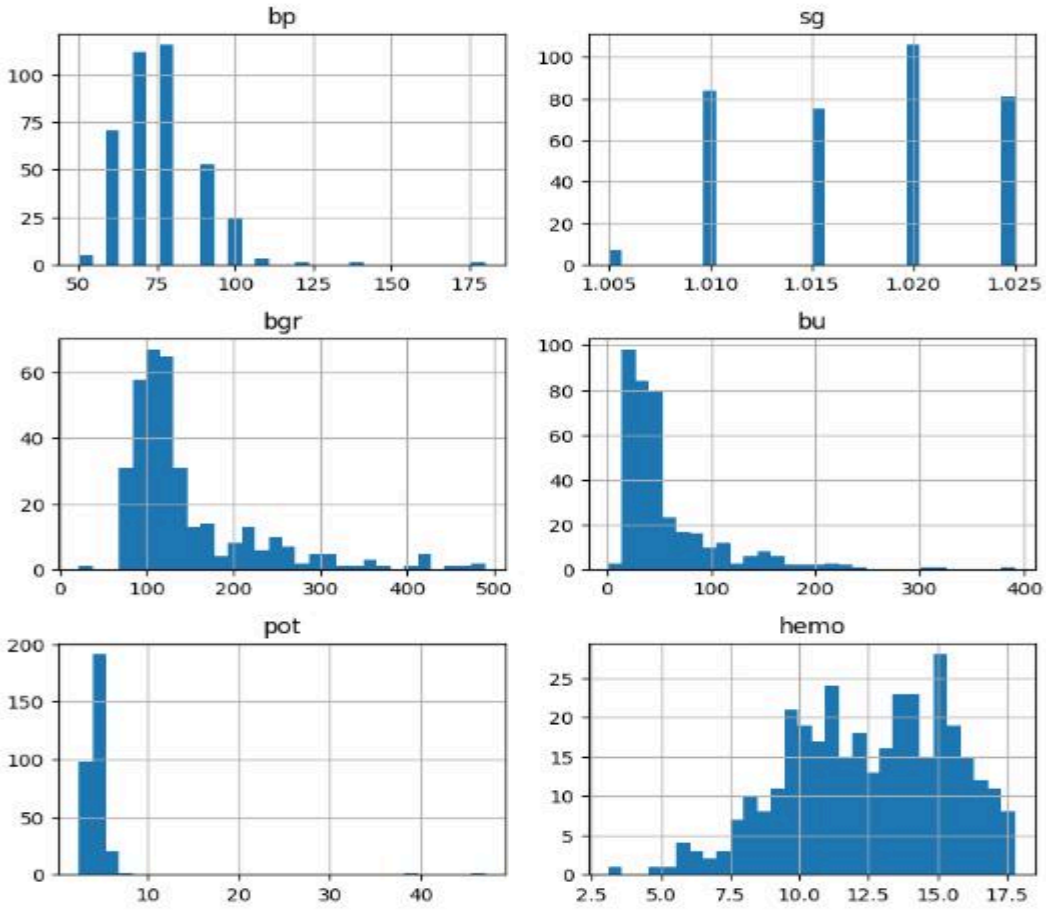
Comparision of Model Analysis

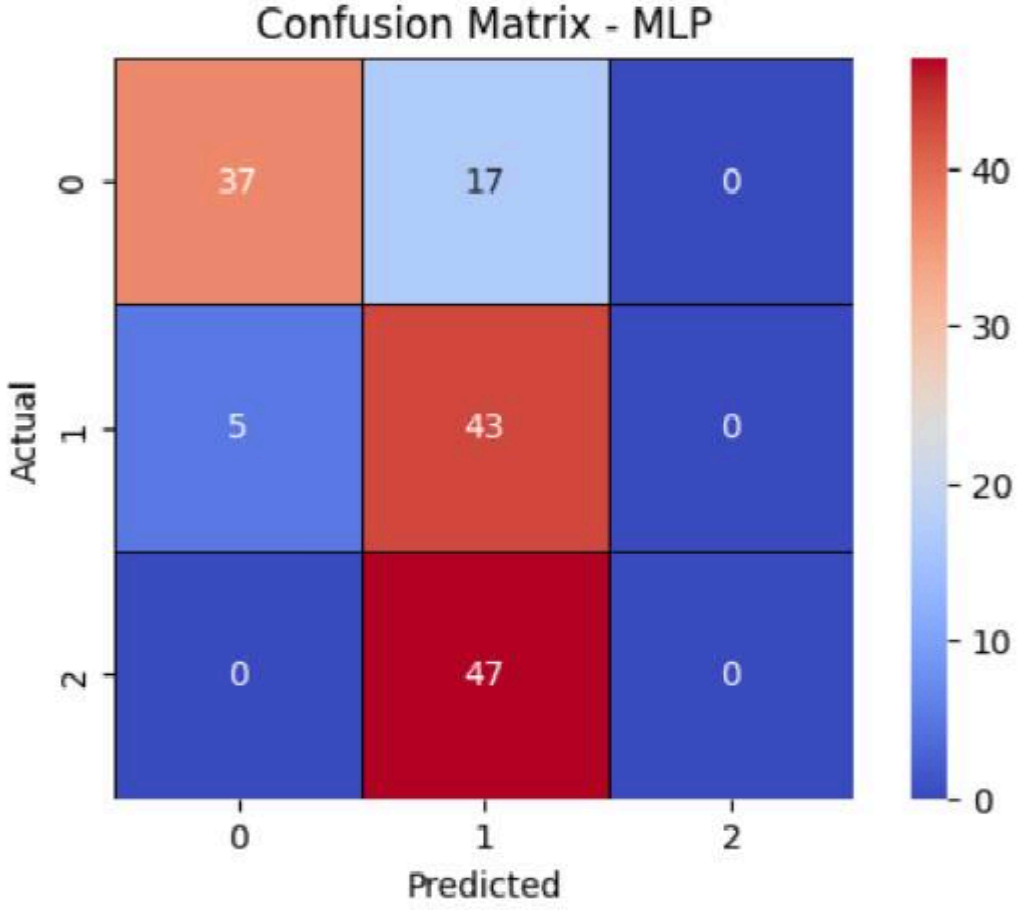
	accuracy	precision	recall	f1	roc_auc
ML logistic	0.986577	0.986854	0.986577	0.986520	1.000000
dt	0.986577	0.986714	1.000000	0.986576	0.989786
rf	1.000000	1.000000	1.000000	1.000000	1.000000
svm	1.000000	1.000000	0.979866	1.000000	1.000000
knn	0.979866	0.981074	0.993289	0.979889	1.000000
xgb	0.993289	0.993411	NaN	0.993283	0.999859
DL MLP	0.536913	NaN	NaN	0.458102	NaN
CNN	0.375839	NaN	NaN	0.300971	NaN
LSTM	0.624161	NaN	NaN	0.532931	NaN
Autoencoder	0.241611	NaN		0.189087	NaN

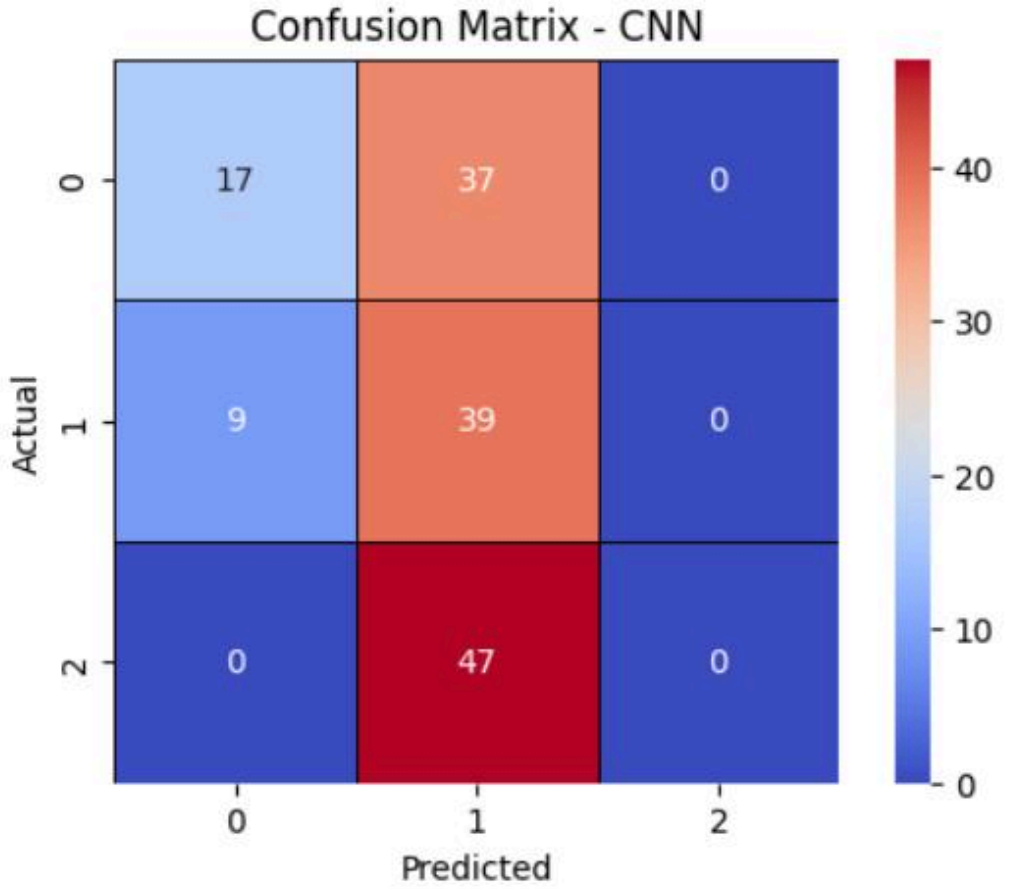
Visualizations

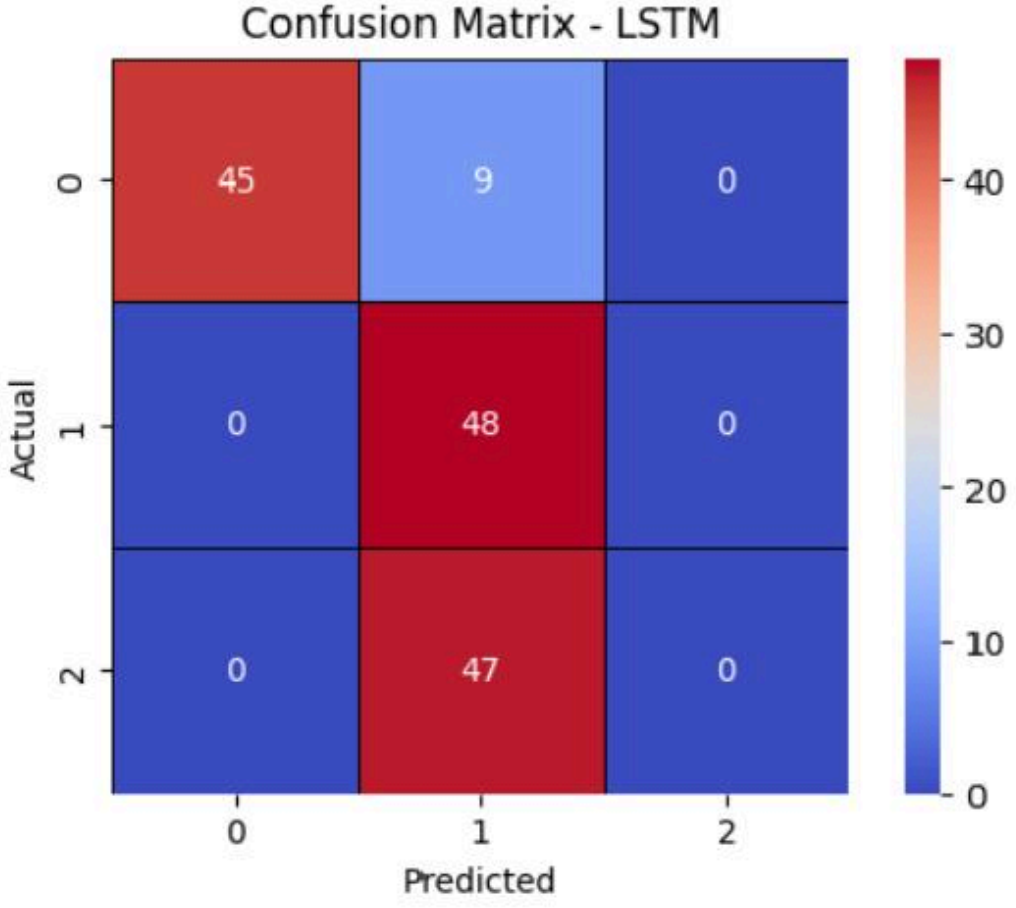


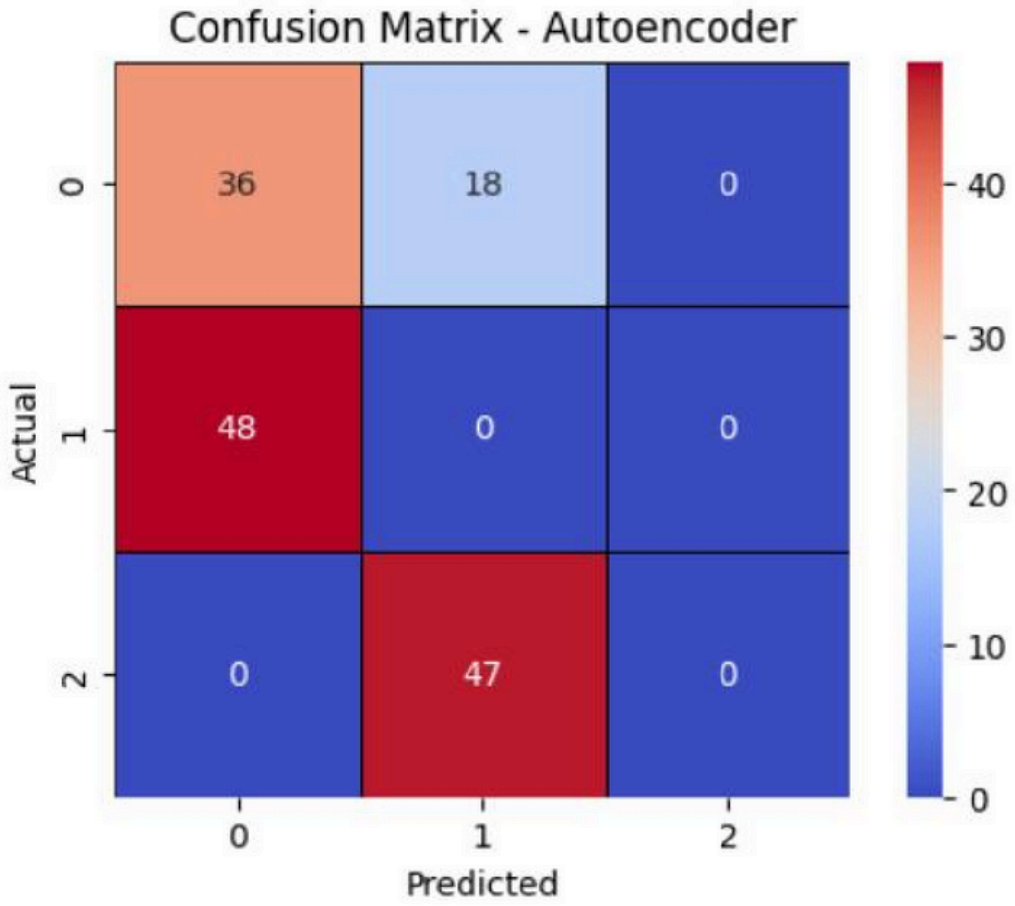


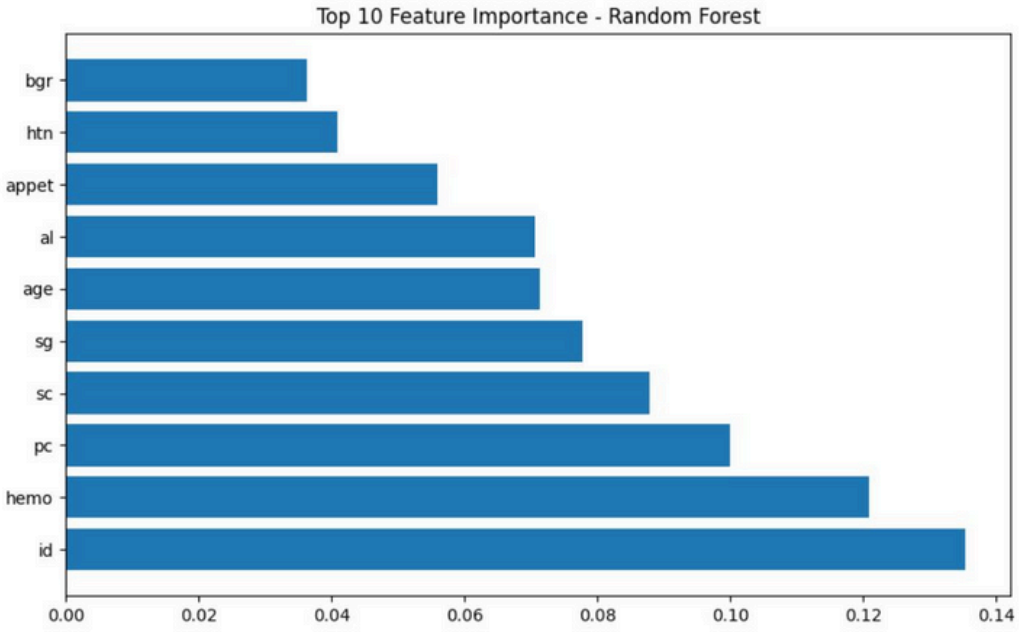

bp
sg
bgr
bu
pot
hemo

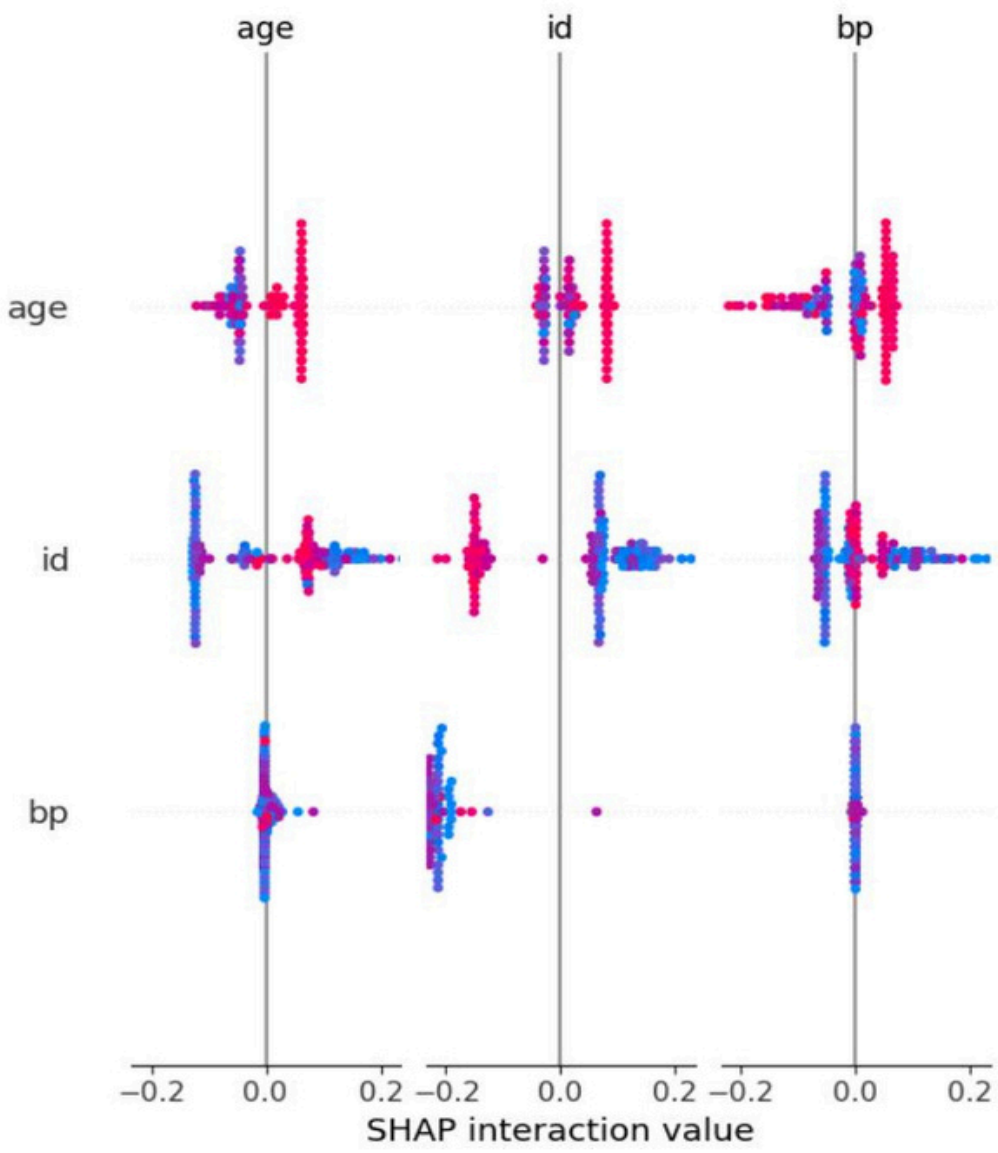

Confusion Matrix - MLP
37
17
0
5
43
0
0
47
0
Actual
Predicted

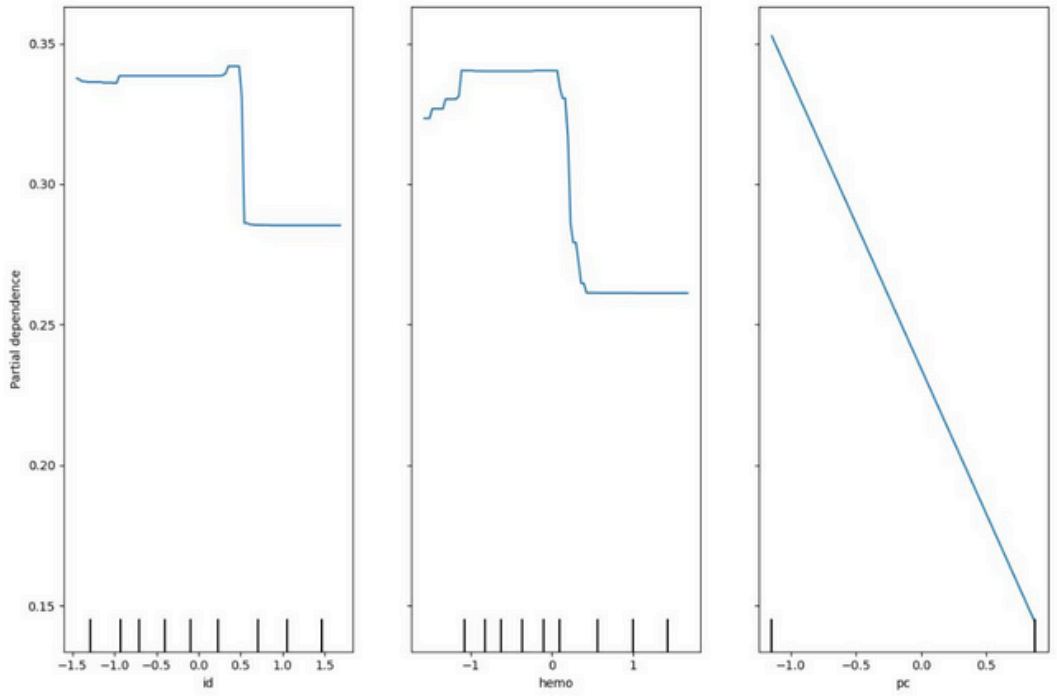

Confusion Matrix - CNN
17
37
0
9
39
0
0
47
0
Actual
Predicted
0
1
2
40
30
20
10
0

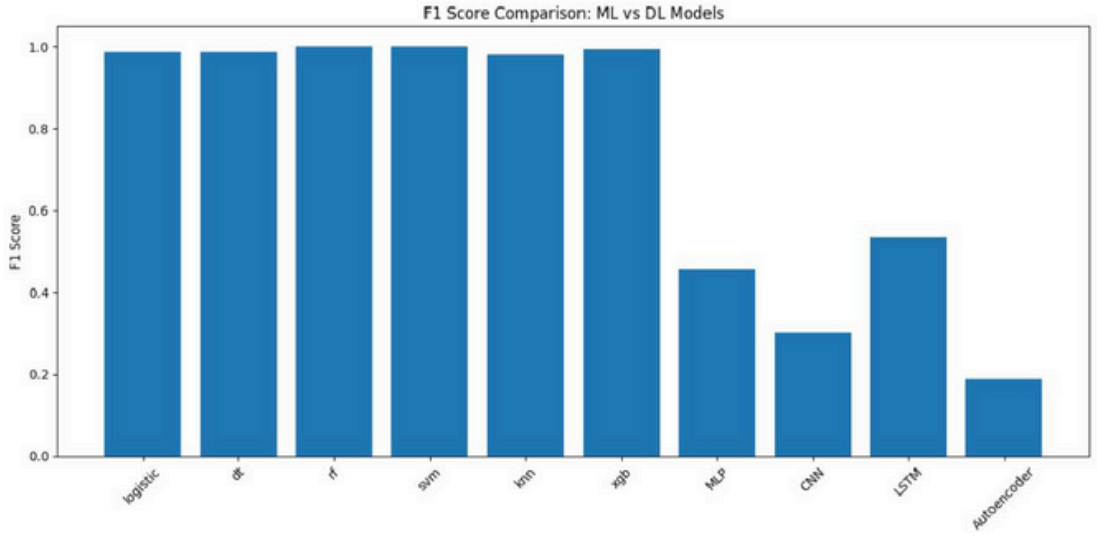

Confusion Matrix - LSTM
45
9
0
0
48
0
0
47
0
Actual
Predicted
0
1
2
40
30
20
10
0


Confusion Matrix - Autoencoder
36
18
0
48
0
0
0
47
0
Actual
0
1
2
Predicted
0
1
2
40
30
20
10
0


Top 10 Feature Importance - Random Forest
bgr
htn
appet
al
age
sg
sc
pc
hemo
id
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14


age
id
bp
age
id
bp
−0.2
0.0
0.2
−0.2
0.0
0.2
−0.2
0.0
0.2
SHAP interaction value





Final recommendation

The XGBoost model is suggested to be deployed based on the elaborate model performance and explainability analysis.

Although the base rates of the Random Forest and ANN models were also 100 percent on the test set, the XGBoost provides a good balance between high

performance and interpretability. XGBoost SHAP analysis can deliver explicit and practical information on feature contributions which is the obvious benefit of a clinical application. The capability to describe the prediction of a model on a particular patient is invaluable and helps allow clinicians to compare the logic of a model with their actual knowledge and the entire clinical history of the patient.

Furthermore, one can conclude that the created solution does not only offer a highly precise Chronic Kidney Disease prediction tool but also it does it in such a way that the solution is transparent enough to be implemented in a real-life healthcare environment. This is a strong and reliable model that can be trusted because of its strong predictive power and interpretability offered by XAI.