DESIGN OF CHRONIC KIDNEY DISEASE PREDICTION MODEL ON IMBALANCED DATA USING MACHINE LEARNING TECHNIQUES

Vinothini A
Assistant Professor, Department of Computer Science and Engineering,
Rajalakshmi Engineering College, Chennai, Tamilnadu, India
vinothini.a@rajalakshmi.edu.in

Baghavathi Priya S
Professor, Department of Information Technology,
Rajalakshmi Engineering College, Chennai, Tamilnadu, India.
baghavathipriya.s@rajalakshmi.edu.in

Abstract - The objective of this paper is to build a CKD prediction model using machine learning techniques that can predict the risk of chronic kidney disease (CKD) in patients with Cardiovascular Disease (CVD) or at high risk of CVD. CVD is associated with worsening of renal functions. But patients with CVD remains often underdiagnosed and undertreated for CKD because mostly the clinical diagnosis and treatment are single organ centered in earlier stages. Machine learning algorithms have been widely used to predict and classify diseases in healthcare. Healthcare data is often imbalanced. In this analysis, the CKD prediction model is built using CVD data with imbalanced distribution of positive and negative cases. The analysis involves three stages: Stage I involves selecting the best model based on performance metrics that support imbalanced class distribution without applying any resampling techniques. Stage II involves oversampling the training data of the minority class using Synthetic Minority Oversampling Technique (SMOTE) and stage III involves randomly under-sampling the training data of the majority class to solve the class imbalance. The experimental results show that the MLP (Multi-Layer Perceptron)-SMOTE model performs better in predicting CKD with a better F-score, recall, precision, G-mean, balanced accuracy and RUC-AUC when compared to other models.

Keywords: Chronic kidney disease; class imbalance; machine learning; sampling techniques.

1. Introduction

CKD is one of the leading causes of morbidity and mortality for individuals with CVD. A precise CKD risk prediction model developed from CVD patient data is critical for secondary prevention of CKD. Artificial Intelligence (AI) is enhancing the astuteness of medical professionals in diagnosis and prognosis in the field of nephrology. Early diagnosis of kidney disease by AI will help the health practitioners to screen potential kidney disease patients according to their risk levels. The availability of a huge volume of data and also with high quality is the greatest challenge in building an accurate and most efficient machine predictive model [Xie *et al*., (2020)].

Various classifiers are deployed to create classification models for classifying the clinical CKD data [Jena *et al*.,(2020)]. Supervised learning algorithms like logistic regression (LR), neural network (NN) and support vector machine (SVM) are used for creating classification models in [Ahmad *et al*., (2020)]. Classification and association rule mining techniques are unified and utilized to paradigm a system for predicting and diagnosing CKD and its roots. Among, naive bayes (NB), decision tree (DT), SVM, k-nearest neighbor (K-NN) and JRip experimented on the medical data, K-NN achieved the highest accuracy. Apriori algorithm is applied to selected attributes of CKD data to extract strong rules based on the lift matrix [Alloghani *et al*., (2020)].

Statistical method like multivariable cox's proportional hazards analysis is applied in [Shamsi *et al*., (2018)] over the high-risk CVD patients to determine the independent risk aspects like older age, history of coronary heart disease (CHD), diabetes mellitus, and smoking associated with developing CKD stages 3 to 5. Machine learning methods like multivariate regression model, classification and regression tree (CART), NB, bagged trees, ada boost and random forest are used in [Yang *et al*.,(2020)] to build a prediction model for CVD disease. The CVD prediction model provided a three-year risk assessment of CVD. The prediction performance becomes unsatisfactory when prediction models are deployment into the local population. A knowledge-enhanced localized risk model is developed by [Mei *et al*.,(2017)] to solve the localization issue. [Niehaus and Clifton , (2016)] proposed an extreme value theory (EVT) that can be applied to better quantify severity and risk in chronic disease.

The increasing research importance on novel machine learning approaches recommends that the modeling of chronic disease will continue to yield valuable discoveries for patients and doctors [Karthikeya and Menakadevi ,(2020)]. Machine learning algorithms perform better when there is an equal number of records collected for all the target classes [Santos *et al*.,(2018)]. But in reality, medical data for disease prediction cannot always be collected with equal distributions for diseased and non-diseased target classes. In the case of binary classification problem, one target class may have many instances than another, thus leading to an imbalance in the dataset [Thabtah *et al*.,(2020)]. This problem is referred to as class imbalance.

The rest of the paper is organized as follows: section 2 elucidates the related work. Section 3 presents the materials and methods. Section 4 describes the results and discussions. Finally, section 5 presents the conclusion and outlines future work.

2. Related Work

In machine learning, the quality and quantity of input data that is used for training the classifiers are very important. Most algorithms perform well when the prior probabilities of the target classes are similar. Data is said to be imbalanced if at least one of the target variable values has a significantly smaller number of instances when compared to the other values. Class imbalance is one of the vital issues in machine learning classification tasks. Machine learning algorithms trained on imbalanced data emphasize exploiting the total accuracy over the entire dataset leading to more attention being paid to the majority class samples [Thabtah *et al*.,(2020)]. Due to this scenario, the minority class samples are poorly projected by the learning model.

[Shuja *et al*.,(2020)] proposed a two-phase classification model to solve the class imbalance problem for predicting type II diabetes. SMOTE was used to rebalance the data. The pre-processed data was then trained using a DT classifier. The DT showed increased performance by reducing class imbalance. Class imbalance causes difficulties for classifiers. [Thabtah *et al*.,(2020)] studied data-driven methods and the algorithm-driven approaches for dealing with class imbalance problems for the autism diagnosis. The approaches on data fine-tune the class proportion in the input data to generate a balanced dataset. In algorithm-driven approaches the classification algorithm is fine- tuned to create a model that learns more from the minority class. Thus, the dataset will remain as imbalanced. In this approach, no changes are made to the input data distribution. [Carrington *et al*., (2020)] proposed a concordant partial area under the receiver operating characteristic (ROC) for measuring the performance of the machine learning algorithms on imbalanced data with low prevalence.

[Sarkar *et al*., (2020)] experimented with SMOTE, borderline SMOTE (BLSMOTE), majority weighted minority oversampling technique (MWMOTE), and k-means SMOTE (KMSMOTE) to handle the class imbalance issue in the prediction of injury severity using machine learning techniques. [Nnamoko and Korkontzelos ,(2020)] tackled class imbalance with SMOTE for predicting diabetes. Experiments were conducted with NB, SVM-radial basis function, C4.5 and repeated incremental pruning to produce error reduction (RIPPER). SMOTE sampled data applied to C4.5 DT produced better results than the other three classifiers.

[Zhang *et al*.,(2020)] developed an active balancing mechanism for the biomedical data under- sampling. Gaussian NB and entropy were used to evaluate sample information and retain valuable samples of the majority class to achieve under-sampling. [Richhariya and Tanveer,(2020)] proposed a reduced universum twin SVM for dealing with class imbalance learning. The model makes a balanced environment for the classification by incorporating prior information from the universum data points.

[Tao *et al*.,(2020)] designed an affinity and class probability based fuzzy SVM (ACFSVM) approach for imbalanced datasets classification tasks. The proposed technique gives more importance to majority class samples with higher affinities and class probabilities ultimately skewing the final classification boundary toward the majority class. But the minority class samples are dispensed with high memberships to promise their importance for the learning. [Dubey *et al*.,(2014)] resolved the problem of imbalance by under-sampling the training neuro imaging data during their work to analyze Alzheimer's disease. [Peng *et al*.(2019)] proposed a trainable under-sampling method that applies evaluation metric optimization into the data sampling procedure. By using such an optimization, the method learns the instances to be discarded and the instances to be preserved.

Previous researches on class imbalance have focused on a few diseases but not on predicting CKD. In this work, a CKD predictive model is built by balancing the imbalanced CVD data by applying SMOTE over minority class and then training with MLP (Multi-layer Perceptron) classifier.

3. Materials and Methods

3.1. *Data source*

Table 1. Imbalanced class distribution of CVD dataset

	Class	
	No	**Yes**
Count	435	56
Percentage	88.6	11.4

The data for analysis is an electronic medical record (EMR) collected by [Shamsi *et al.*, (2018)] during their research on CKD. The CVD dataset has 23 features of CVD or at- risk CVD patient and 491 instances. The target class “EVENTCKD35” has two values “Yes” and “No”. CKD cases are represented as “Yes” and normal cases are represented as “No”. Table 1 shows the number of positive cases for CKD in CVD patients’ data (class 1) is lesser than the number of negative cases for CKD (class 0). This makes it clear that the dataset that is taken for experimentation has class imbalance problem.

3.2. *Proposed CKD prediction experimental framework*

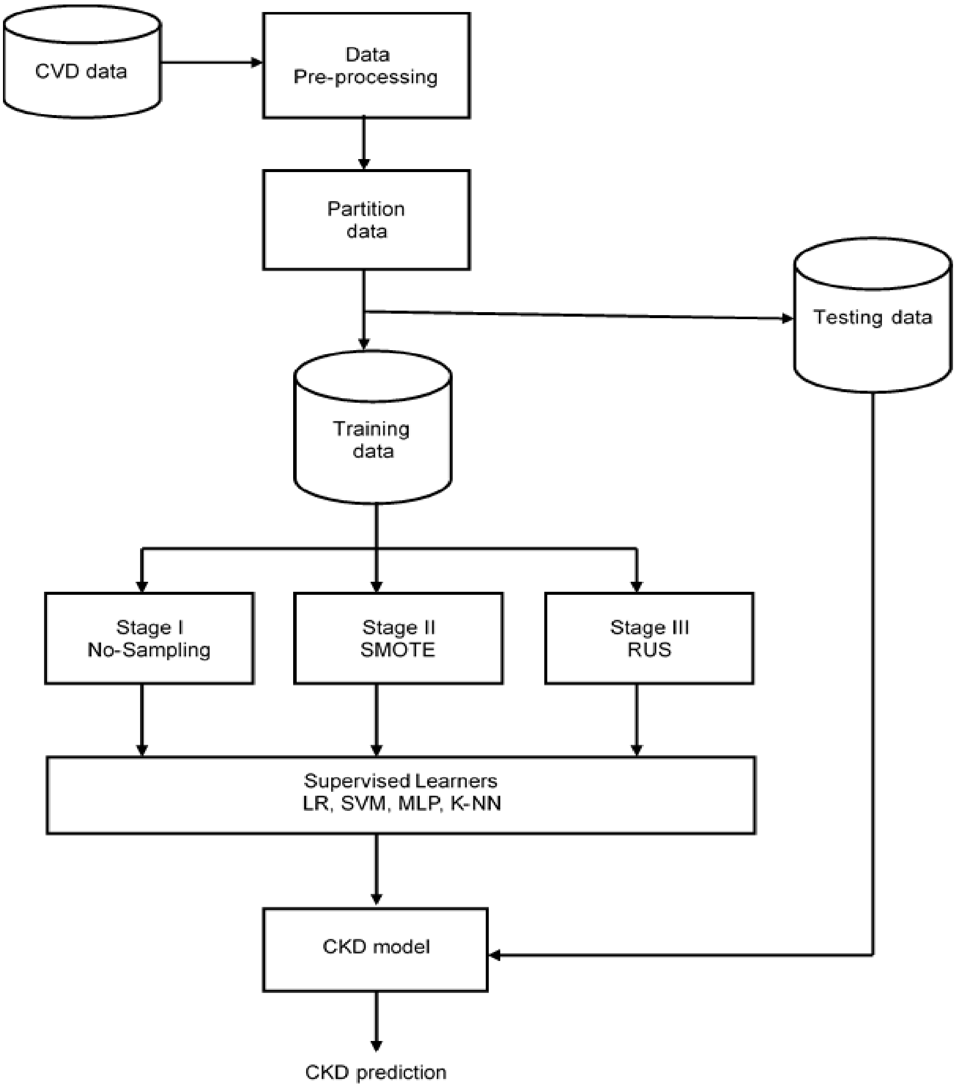


Fig. 1. Proposed CKD prediction experimental framework

The proposed CKD prediction experimental framework on class imbalanced CVD data is shown in Figure 1. The proposed framework involves steps like data pre-processing, solving class imbalance problem using sampling techniques, building the model using supervised learners and selecting the best models that outperforms others in terms of performance metrics. The framework also involves in applying the imbalanced data in building the model. Results obtained on predictions using no sampling and sampling techniques are compared.

3*.2.1. Data pre-processing*

The dataset is pre-processed before feeding it to a classification algorithm. The features in the CVD dataset consist of patient history details like gender, age, age-categories (split into three age groups) and the presence or absence of health conditions like diabetes mellitus (DM), CHD, vascular disease, smoking status, hypertension (HTN), dyslipidemia (DLD), obesity. It also includes the history of intake of medicine for DLD, DM, HTN and angiotensin-converting enzyme inhibitors/angiotensin II receptor blockers (ACEI/ARB). Laboratory values includes cholesterol, triglycerides, HgbA1C (glycosylated Hemoglobin, type A1C), Creatinine, sBP (Systolic blood pressure), dBP (Diastolic blood pressure), eGFR (estimated Glomerular Filtration Rate), BMI (Body mass index). Further, it includes the patient observation time in months and the target class label 'EventCKD35' [Shamsi *et al.*, (2018)]. The input features age-categories and age represent the same information about age. So, the age-categories feature is removed. Also, the feature 'eGFR' calculated using standard formulas may determine CKD directly [6]. This may create a hindrance to learn about the other features that may contribute to detect CKD. So, it was removed. The dataset contains a few missing values for the features 'triglycerides' and 'HgbA1C'. Since the count of missing values is very less, the instance with missing values is ignored. After removing missing values, 460 instances were representing the medical records of CVD patients.

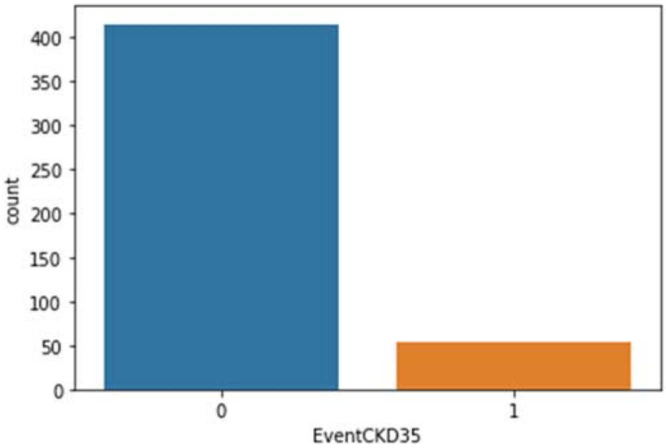


Fig. 2. Class distribution

The class distribution is shown in Figure 2. There are 415 class 0 records and 54 class 1 records. The dataset has 21 input features and one target feature named as 'EventCKD35'. Among the 21 input features, 9 features are numerical in nature and 12 features are categorical with values "yes" and "no". The values for categorical features are encoded to numerical 1 or 0 using label and one hot encoding technique. The range in the distribution of numerical features varies widely. To normalize the values of features, all numerical feature is scaled between values -1 and 1 using the min-max scaler as shown in Figure 3. After data pre-processing, the imbalanced data is divided as 70% training data and 30% testing data.

3*.2.2. Class imbalance*

Imbalanced classification problems represent a classification model created with data in which the distribution of class values across the classes is not equal [Belarouci and Chikh, (2017)]. A classification problem gets skewed because of the nature of class imbalance [Tyagi and Mittal,(2020)]. In this analysis 88.6% of the class values belong to the majority class "No" and 11.4% of the class values belong to the minority class "Yes". But the prediction model is intended to create a model to predict the CKD patients from the CVD or high-risk CVD patient's data. In this case, classification accuracy (A) can mislead to select the best performing model. Techniques to select the best model for data with class imbalance are: Choosing the performance metrics those that focus on the minority class, oversampling the minority class using SMOTE to rebalance the class, under-sampling the majority class to rebalance the class and selecting classification algorithms such as those that penalize misclassification errors differently [Zhao *et al.*,(2018)]. The classification algorithms such as LR, SVM, MLP and K-NN are used for creating classification model.

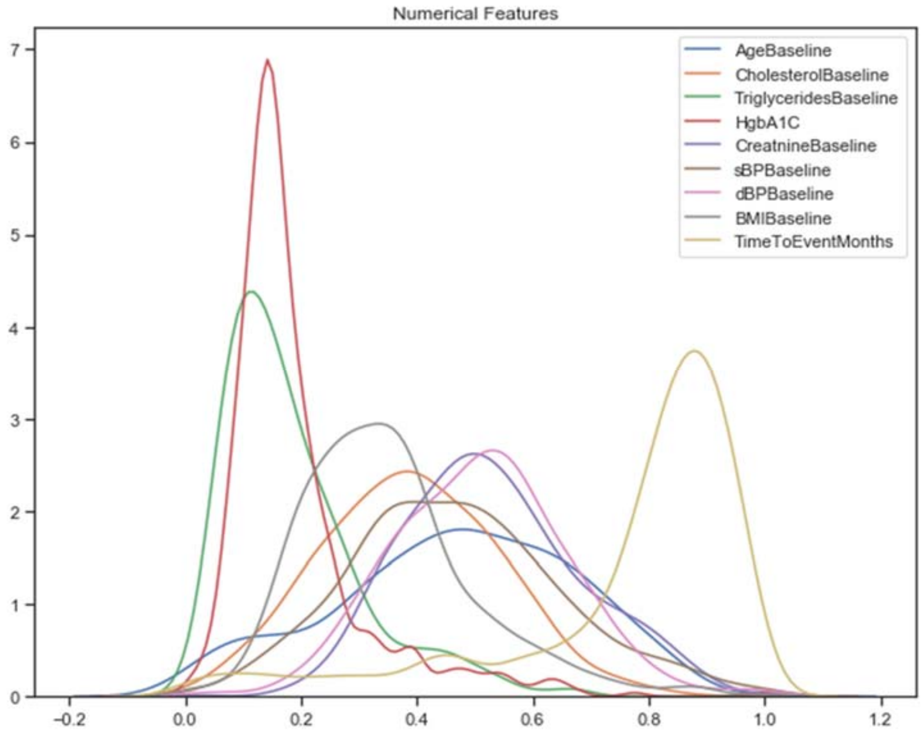


Fig. 3. Numerical features distribution

3.*2.3. Oversampling using SMOTE*

In this oversampling approach, the minority class is over-sampled by creating "synthetic" instances rather than by over-sampling with replacement [Chawla *et a*l., (2001)]. Oversampling encompasses accumulating samples to the minority class in an exertion to reduce the skew in the class distribution [Pan *et al*.,(2019)]. The minority class "Yes" is over-sampled, which means the number of samples is increased. SMOTE iterates through the existing minority instance. At each iteration, one of the 'X' closest minority class neighbors are chosen. A new minority instance is synthesized at some point between the minority instance and that neighbor. Synthetic examples are inserted along the line segments joining any of the X minority class nearest neighbor or all of the X minority class nearest neighbors. Depending upon the amount of over-sampling, N neighbors are chosen. Synthetic samples are created by taking the difference between the feature vector under consideration and its nearest neighbor. The difference is then multiplied by a random number between 0 and 1. The obtained result is then added to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features [Blagus and Lusa,(2013)].

3.*2.4. Under sampling using random under sampling*

Under-sampling techniques eliminate instances from the training dataset that belong to the majority class in order to reduce the skew in class distribution [Mohammed *et al*. ,(2020)]. The instances can be removed in the ration 1:1, 1:2, 1:100 or any ratio according to requirements. There are different techniques in under-sampling such as random majority under sampling, near miss, cluster centroid and Tomek link [Brownlee ,(2020)]. In this work, random under sampling (RUS) is used. The samples of the majority class "No" of the training data are randomly removed such that a balanced 1:1 class distribution is created. Classifiers are trained on this balanced dataset.

3.*2.5. Stages of Analysis*

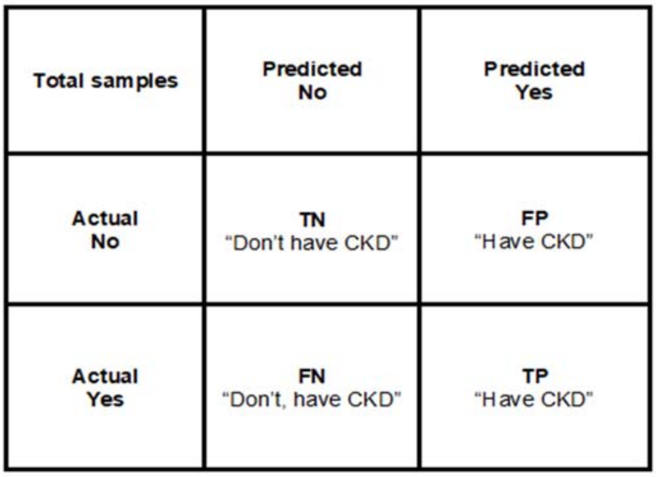
To build an efficient CKD prediction model, the experiment is done in three stage's as shown in Figure 1.

- Stage I: Building a classification model with the imbalanced data (No Sampling).
- Stage II: Building a classification model by oversampling the minority class using SMOTE to rebalance the class.
- Stage III: Building a classification model by under-sampling the majority class using RUS to rebalance the class.

In stage I the training data is fed to classifiers. The classification algorithms such as LR, K-NN, SVM, MLP were fit on the training data. The training data was validated with 10-fold cross-validation. The models obtained were evaluated using various performance metrics that support the minority class 1 to choose the best model. Stage II aims in building a classification model by oversampling the minority class using SMOTE to rebalance the class. The minority class "1" is oversampled using SMOTE to balance the class distribution. The training data is fed to SMOTE sampler to oversample the minority class 1. Class 1 is oversampled with instances equal to the number of majority class 0. After this pre-processing stage the data becomes balanced. The balanced training data is fed to classifiers and classification model is created. The test data is then fed to the models to predict the risk of CKD.

Stage III intends to build a classification model by under sampling the majority class using RUS to rebalance the class distributions. The majority class "0" in the training data is under-sampled randomly by deleting instances in the class"0". By deleting the instances of majority class, the information that classifiers can learn is lost [Koziarski, (2020)]. But now the class becomes balanced. With equal distributions in both the class, the classifiers build the balanced model. The imbalanced test data is fed to the models to predict CKD.

4. Results and Discussion



Total samples	Predicted No	Predicted Yes
Actual No	TN "Don't have CKD"	FP "Have CKD"
Actual Yes	FN "Don't, have CKD"	TP "Have CKD"

Fig. 4. Confusion matrix

Performance metrics that focus on the minority class are sensitivity or recall (R), precision (P), F-score, balanced accuracy (BA) and geometric mean (G-Mean). The confusion matrix shown in Figure.4 is a valuable tool in analyzing the predicted and actual values and supports to measure the performance of the predictive model [He and Garcia, (2009)]. True positive (TP) gives the number of observations correctly predicted as CKD, false positives (FP) gives the number of observations that are incorrectly predicted as CKD which are not CKD. True negative (TN) tells the number of observations predicted correctly as not having CKD and false negative (FN) gives the number of observations incorrectly predicted as not having CKD.

$$R = \frac{TP}{TP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

R is the True Positive Rate (TPR). It gives the information about how well the positive class "Yes" is predicted as shown in Eq. (1). P states the information about the fraction of observations that are really positive out of all the observations that are predicted as positive as in Eq. (2). F-score in Eq. (3) gives the balance between P and R. True negative rate (TNR) in Eq. (4) can be measured to know how well TN is predicted.

$$\mathrm{F-score} = \frac{\mathrm{2PR}}{(\mathrm{P} + \mathrm{R})} \tag{3}$$

$$TNR = \frac{TN}{(TN + FP)} \tag{4}$$

$$BA = \frac{(TPR + TNR)}{2} \tag{5}$$

$$G - mean = \sqrt{TPR \times TNR} \tag{6}$$

The BA metric in Eq. (5) is a more suitable metric to measure the performance of classifiers on imbalanced data. [Luquea *et al*.,(2019)]. The G-mean in Eq. (6). shows a balance in classification performance in terms of R and TNR [Wang *et al*., (2018)]. The performance of classifiers in phase I analysis is evaluated. Table 2 shows that the classifiers LR, SVM, MLP and K-NN perform with A>=90%. Among the four classification algorithms, MLP has the highest A - 93%. But R and F-score are very less. It is clear that the accuracy is biased by the majority class value 0. This is because the CVD data collected for predicting CKD suffers from severe class imbalance problem. If A alone be used as a metric to select the best model, it can mislead the classification task. So, the focus must be on the other performance metrics that can be used to evaluate the CVD data which has class imbalance issues. The evaluation metrics that can be used on the balanced data treat all classes with equal importance. But in imbalanced classification task classification errors with the minority class are more important than those with the majority class.

Table 2. Performance of classifiers on imbalanced training data

Model	A	P	R	F-score	ROC- AUC
LR	0.90	0.67	0.25	0.36	0.62
SVM	0.91	1.00	0.25	0.40	0.62
MLP	0.91	0.67	0.50	0.57	0.73
K-NN	0.87	1.00	0.12	0.22	0.56

Table 3. Predictions on imbalanced testing data (No sampling)

Model	A	P	R	F-score	ROC- AUC
LR	0.90	0.67	0.25	0.36	0.62
SVM	0.91	1.00	0.25	0.40	0.62
MLP	0.91	0.67	0.50	0.57	0.73
K-NN	0.87	1.00	0.12	0.22	0.56

Table 4. Comparison of predictions of imbalanced models

Model	TN	FP	FN	TP	TPR	FPR	TNR	FNR	BA	G-mean
LR	123	2	12	4	0.25	0.02	0.98	0.75	0.62	0.50
SVM	125	0	12	4	0.25	0.00	1.00	0.75	0.63	0.50
MLP	121	4	8	8	0.50	0.03	0.97	0.50	0.73	0.70
K-NN	125	0	14	2	0.12	0.00	1.00	0.88	0.56	0.35

Table 5. Performance of classifiers on SMOTE balanced training data

Model	A	P	R	F-score	ROC- AUC
LR-SMOTE	0.92	0.75	0.93	0.51	0.91
SVM-SMOTE	0.93	0.91	0.96	0.94	0.94
MLP -SMOTE	0.91	0.90	0.98	0.94	0.93
K-NN-SMOTE	0.90	0.63	0.97	0.38	0.77

Table 6. Performance of SMOTE models on testing data

Model	A	P	R	F-score	ROC- AUC
LR-SMOTE	0.87	0.43	0.56	0.49	0.73
SVM-SMOTE	0.89	0.53	0.62	0.57	0.78
MLP -SMOTE	0.93	0.64	0.56	0.60	0.76
K-NN-SMOTE	0.81	0.31	0.56	0.40	0.70

Table 7. Comparison of predictions of SMOTE models

Model	**TN**	**FP**	**FN**	**TP**	**TPR**	**FPR**	**TNR**	**FNR**	**BA**	**G-mean**
LR-SMOTE	113	12	7	9	0.56	0.10	0.90	0.44	0.73	0.71
SVM-SMOTE	116	9	6	10	0.62	0.07	0.93	0.38	0.76	0.76
MLP -SMOTE	120	5	7	9	0.56	0.04	0.96	0.44	0.76	0.73
K-NN-SMOTE	105	20	7	9	0.56	0.16	0.84	0.44	0.70	0.69

Table 8. Performance of classifiers on RUS balanced training data

Model	**A**	**P**	**R**	**F-score**	**ROC- AUC**
LR-RUS	0.75	0.78	0.77	0.74	0.88
SVM-RUS	0.77	0.76	0.82	0.78	0.83
MLP-RUS	0.79	0.82	0.81	0.78	0.90
K-NN-RUS	0.73	0.78	0.66	0.71	0.77

Table 9. Performance of RUS model on testing data

Model	**A**	**P**	**R**	**F-score**	**ROC- AUC**
LR-RUS	0.84	0.38	0.62	0.48	0.75
SVM-RUS	0.83	0.36	0.75	0.49	0.79
MLP-RUS	0.85	0.42	0.81	0.55	0.83
K-NN-RUS	0.83	0.34	0.75	0.47	0.78

Table 10. Comparison of predictions of RUS models

Model	**TN**	**FP**	**FN**	**TP**	**TPR**	**FPR**	**TNR**	**FNR**	**BA**	**G-mean**
LR-RUS	109	16	6	10	0.62	0.13	0.87	0.38	0.75	0.87
SVM-RUS	106	19	5	11	0.69	0.15	0.85	0.31	0.77	0.88
MLP-RUS	107	18	3	13	0.81	0.14	0.86	0.19	0.84	0.92
K-NN-RUS	107	18	6	10	0.62	0.14	0.86	0.38	0.74	0.86

Table 11. Comparison of imbalanced and balanced models

Classifier	**Imbalanced model**			**SMOTE model**			**RUS model**		
	F-score	G-mean	BA	F-score	G-mean	BA	F-score	G-mean	BA
LR	0.36	0.50	0.62	0.49	0.71	0.73	0.48	0.87	0.75
SVM	0.40	0.50	0.63	0.57	**0.76**	**0.76**	0.49	0.88	0.77
MLP	**0.57**	**0.70**	**0.73**	**0.60**	0.73	**0.76**	**0.55**	**0.92**	**0.84**
K-NN	0.22	0.35	0.56	0.40	0.69	0.70	0.47	0.86	0.74

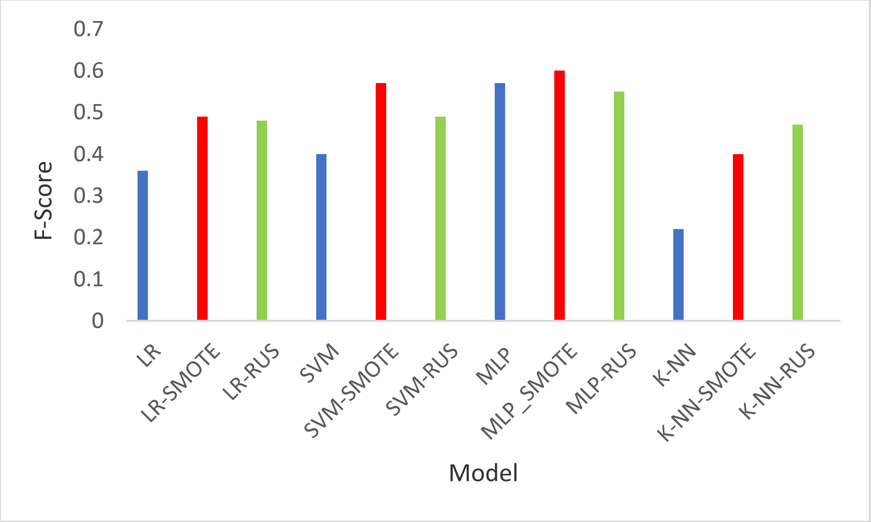


Fig.5. F-score of imbalanced and balanced models

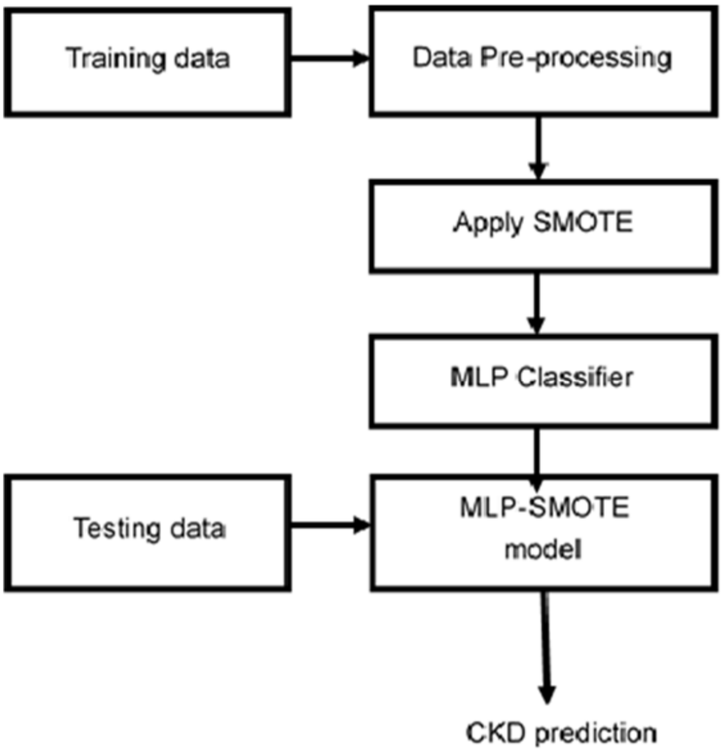


Fig. 6. Proposed CKD prediction model

The challenge is to choose the metric that focuses on the minority class where we dearth observations required to train an effectual classification model. P, R, F-score are the metrics that can measure the effectiveness of the model created when there is a skew in the class distribution. Based on this, MLP produces the highest R of 79%, P of 69% and F-score of 70%. The classification model build is fed with the testing data to observe how it performs on the test data. Table 3 shows that the MLP classifier has a better R and F-score when compared with the other models. MLP and SVM performed better than LR and KNN classifiers by achieving 69% P, 79% R and 70% F-score. SVM and LR attained 25% R. MLP and LR reached 67% P. But SVM has the highest P of 100% with a low R of 25%. The reason for the very low R rate in all classifiers is due to the skew distribution of the class. MLP performed better than other classifiers in terms of area under the receiver operating characteristic curve (ROC-AUC) in phase I. Table 4 shows that the MLP also achieved the highest BA -73% and G-mean – 70 % when compared with the other three models.

In phase II, classifiers are applied to the SMOTE oversampled CVD data. The balanced data is 10-fold cross-validated. In the training phase SVM -SMOTE model produced the highest A- 93%, P-91%, F score - 94% and ROC-AUC 94%. Table 5 shows the performance of the classifiers on the oversampled training data. In all the four classifiers R rate has increased when compared with model performance on imbalanced data in Table 2. This is because the data is now balanced. MLP-SMOTE achieved the highest R rate. These models were applied to the test data and evaluated as shown in Table 6. The test data is not balanced. On the test data, MLP attained better performance than the other classifiers in terms of P and F-score. Table 7 shows that SVM -SMOTE produced the highest G-mean of 76%. But MLP-SMOTE has better TNR and BA. In phase III, classifiers were trained with under- sampled data. Table 8 shows that the accuracy of all the RUS models on training data has decreased when compared with no-sampling and SMOTE models. But the MLP-RUS performs better than other RUS models in terms of A, P, R, F-score and ROC- AUC as shown in Table 9.

Table 10 shows that the G-mean of MLP-RUS has increased which shows that the performance is validated with equal importance to both TPR and TNR. The BA of MLP-RUS is also higher when compared with all other models. MLP classifier performs better when oversampled or under-sampled when compared with other classifiers as shown in Table 11. But when the class 0 data is under-sampled for sake of balancing with class 1, the classifiers miss the opportunity to learn from more data. The F-score of MLP- RUS model has decreased when compared with imbalanced models as shown in Figure. 5. But in MLP-SMOTE model the F-score has increased when compared with imbalanced models.

The proposed methodology offers MLP - SMOTE as the better classifier for predicting CKD from imbalanced CVD as illustrated in Figure. 6. The MLP model is trained with SMOTE balanced data to create an MLP-SMOTE model. The created model is fed with the testing data which is not balanced. By oversampling the imbalanced data, the MLP classifier performance increases. The MLP-SMOTE predicts whether the patient has CKD or not. This prediction result may help the health care practitioners and the patients for the earlier identification of CKD from CVD data.

5. Conclusion

In this work, the CKD prediction model is built using imbalanced CVD data. Initially, the models LR, SVM, MLP and KNN build on imbalanced data, performed with good accuracy in predicting the CVD test data. But R and F-score were low. The low values are due to the fact that the number of CKD positive cases is too low. On the imbalanced data, the MLP model performed better than other models. The CVD training data is then balanced by applying resampling techniques like SMOTE and RUS. On the test data, MLP-SMOTE performed better with the highest and increased F-score when compared with other models. The proposed MLP-SMOTE model can predict CKD better by solving the imbalanced distribution of CVD data. This can help the medical practitioners and patients for the early prediction of CKD and save a life. In the future, the model can be further tuned by applying feature selection methods to increase the performance in prediction.

References

[1] Ahmad, A.; Hassan, N.; Belal, M. (2020). Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease. Journal of Information & Knowledge Management, 19(1), 1-17.

[2] Alloghani ,M.; Al-Jumeily,D.; Hussain, A,; Liatsis ,P.; Aljaaf ,A.J. (2020). Performance-Based Prediction of Chronic Kidney Disease Using Machine Learning for High-Risk Cardiovascular Disease Patients. In Nature-Inspired Computation in Data Mining and Machine Learning. Studies in Computational Intelligence (pp. 187-206). Springer, Cham.

[3] Belarouci, S.; Chikh, M A. (2017). Medical imbalanced data classification. Advances in Science, Technology and Engineering Systems Journal, 2(3), 116-124.

[4] Blagus, R.; Lusa,, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14, 106. https://doi.org/10.1186/1471-2105-14-106.

[5] Brownlee, J. (2020). Undersampling Algorithms for Imbalanced Classification. https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification.

[6] Carrington, A M,.; Fieguth, P W.; Qazi, H. (January 2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Medical Informatics and Decision Making, 20(4), 1-12.

[7] Chawla, N.V.;,Bowyer, K.W,.;Hall,L.O.; Kegelmeyer, W. P. (2001). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

[8] Dubey, R.; Zhou, J.; Wang, Y.; Thompson, P.M.; Ye, J. (2014). Alzheimer's Disease Neuroimaging Initiative. Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study. Neuroimage, 87, 220-241.

[9] Florkowski, C. M.; Harris, J. C. (2011). Methods of Estimating GFR – Different Equations Including CKD-EPI. The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists, 32(2), 75-79.

[10] He, H.; Garcia,E.A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[11] Jena, L.; Nayak, S.; Swain, R. (2020). Chronic Disease Risk (CDR) Prediction in Biomedical Data Using Machine Learning Approach. Advances in Intelligent Computing and Communication. Lecture Notes in Networks and Systems (pp. 232-239). Singapore: Springer Nature.

[12] Karthikeyan, H.; Menakadevi, T. (2020). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-019-01652-0

[13] Koziarski, M. (2020). Radial-Based Undersampling for imbalanced data classification. Pattern Recognition, 102, 11. https://doi.org/10.1016/j.patcog.2020.10726.

[14] Luquea, A.; Carrasco, A.; Martin, A.; Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216-231.

[15] Mei, J.; Xia, E.; Li, X.; Xie, G. (2017). Developing Knowledge-enhanced Chronic Disease Risk Prediction Models from Regional EHR Repositories. https://arXiv:1707.09706 [cs.AI]

[16] Mohammed, R.; Rawashdeh, J.; Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 11th International Conference on Information and Communication Systems (pp. 243-248). Jordan: IEEE.

[17] Niehaus, K E.; Clifton, D A. (2016). Machine learning for chronic disease. In Machine Learning for Healthcare Technologies (p. 296). IET. Retrieved from https://pdfs.semanticscholar.org/38ec/f47227e0838fb981bfec3ca9837cf459eda0.pdf

[18] Nnamoko, N.; Korkontzelos, I. (2020). Efficient treatment of outliers and class imbalance for diabetes prediction imbalance. Artificial Intelligence In Medicine, 104, 101815.

[19] Pan, T.; Zhao, J.; Wu, W.; Yang, J. (2019). Learning Imbalanced Datasets Based on SMOTE and Gaussian Distribution. Information Sciences, 512, 1214-1233. https://doi.org/10.1016/j.ins.2019.10.048.

[20] Peng, M.; Qi, Z.; Xiaoyu, X.; Tao, G.; Xuanjing ,H.; Yu-Gang, J.; Keyu, D.; Zhigang, C. (2019). Trainable Undersampling for Class-Imbalance Learning. AAAI, (pp. 4707-4714).

[21] Richhariya, B.; Tanveer, M. (June 2020). A reduced universum twin support vector machine for class imbalance learning. Pattern Recognition, 102, 107150.

[22] Santos. M.S.; Soares, J. P.; Abreu, P.H.; Araujo,H.; Santos, J. (November. 2018). Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. IEEE Computational Intelligence Magazine, 13, pp. 59-76.

[23] Sarkar,S.; Pramanik, A.; Maiti, J.; Reniers, G. L. (2020). Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. Safety Science, 125, 104616.

[24] Shamsi,A. I.; Regmi ,D.; Govender, R. D. (2018). Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: A population-based study. PLoS ONE, 13(6). https://doi.org/10.1371/journal.pone.0199920

[25] Shuja, M.; Zaman,M.; Mittal, S. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. Advances in Computing and Intelligent Systems. Algorithms for Intelligent Systems (pp. 195-211). Singapore: Springer.

[26] Tao,X .; Li, Q.; Ren,C.; Guo, W.; He, Q.; Liu, R.; Zou, J. (February 2020). Affinity and class probability-based fuzzy support vector machine for imbalanced data sets. Neural Networks, 122, 289-307.

[27] Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalvesv ,H. (March 2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441.

[28] Tyagi, S.; Mittal, S. (2020). Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning. In K. A. Singh P., Lecture Notes in Electrical Engineering (pp. 209-221). Springer, Cham.

[29] Wang, S.; Minku, L.L.; Yao, X. (2018). A Systematic Study of Online Class Imbalance Learning with Concept Drift. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 29(10), 4802 - 4821.

[30] Xie, G.; Chen, T.; Li, Y.; Chen, T.; Li, X.; Liu, Z. (2020, Jan). Artificial Intelligence in Nephrology: How Can Artificial Intelligence Augment Nephrologists Intelligence? Kidney Diseases, 6(1), 1–6.

[31] Yang, L.; Wu, H.; Jin, X. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Nature, 10, 5245. https://doi.org/10.1038/s41598-020-62133-5.

[32] Zhang,H.; Pirbhulal, S.; Wu,W.; Hugo, V. (March 2020). Active Balancing Mechanism for Imbalanced Medical Data in Deep Learning–Based Classification Models. ACM Trans. Multimedia Comput. Commun. Appl, 16(1), 15.

[33] Zhao,Y,.; Wong, Z. S .; Tsui, K. L. (2018). A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. Journal of Healthcare Engineering, 11. https://doi.org/10.1155/2018/6275435