Analysis of Linear and Non-Linear Classifiers in Imbalanced Data to Predict Diabetes Induced Complications

Aniqa Zaida Khanom[1], Sheikh Mastura Farzana[1], Tahsinur Rahman[1], Sharowar Md. Shahriar Khan[2], and Dr. Md. Ashraful Alam[2]

[1] BRAC University, Dhaka, Bangladesh.
{azkhanom, mastura.farzana,tahsinurrahman5}@gmail.com
[2] Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh.
{sharowar.khan,ashraful.alam}@bracu.ac.bd

Abstract. This paper presents a comparison of linear and non-linear classifiers in predicting health complications of the Kidney and Heart induced by Diabetes Mellitus based on an imbalanced dataset. Over time Diabetes damages various organs in the body- primarily Kidney, Eyes, Heart, and Brain. The onset of these complications can be hard to prevent unless a person is monitored closely. This proposed model uses a time series data of one year that contains 164 features of 779 T2DM patients to predict the risk of Nephropathy and Cardiovascular disease. Methods such as Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Tree, and Random Forest have been used to predict the probability of developing the complications. Random Forest produces the best results with 85 trees for Nephropathy with F1 score of 0.75. Logistic Regression without oversampling gives the best results for Cardiovascular disease; the F1 score is 0.54 when C is 0.3.

Keywords: Imbalanced dataset · Complications prediction · Logistic Regression · Random Forest · Oversampling.

1 Introduction

Clinical databases store large amounts of information about patients and their medical conditions that can be used to discover relationships and patterns among clinical and pathological data using data mining techniques [1]. These can be used for early diagnosis by understanding the progression and features of the disease. In most clinical databases, disease cases are fairly rare as compared with the healthy populations, hence creating an imbalance. Diabetes Mellitus is one such example of a health condition. Several Machine Learning based models exist that deal with Diabetes Mellitus [2]. However, most of these systems only predict the probability of a person having Diabetes shortly. According to IDF Atlas published in 2017, there are around 424.9 million Diabetes patients around the world aged from 20-79 years, of whom 95% suffer from Type 2 Diabetes

Mellitus (T2DM). It is predicted that the number will increase to 628.6 million by 2045 [3]. Diabetes Mellitus can induce other complications like Nephropathy, Cardiovascular disease, Retinopathy, and Diabetic Foot disease [4]. In 2017 alone, 4 million people died all around the world due to Diabetes related complications. This paper proposes a prediction model built using linear and non-linear classifiers that can predict the probability of Nephropathy and Cardiovascular disease onset in T2DM patients. An embedded pipeline has been applied upon an imbalanced dataset containing pathological test results of 779 patients, exploring the effects of imbalance of data in prediction models. Additionally, it is different from the conventional Diabetes predicting systems since it emphasizes on predicting Diabetes related complications. There is a scope to introduce a complete system that can correctly predict the onset of complications caused by T2DM using Machine Learning techniques.

2 Related Work

Researchers have found various new aspects in past years discussing the factors which assist Diabetes development and its impact on health that causes different types of complications [5]. The recent progress in Deep Learning systems has made it possible to develop a system that diagnoses retinal diseases [6]. AI and Machine Learning are assets that can help doctors overcome their limitations by improving the prediction and diagnosis of conditions [7]. In July 2018, an AI system beat 15 Chinese doctors in a tumor detection competition which is the latest among several similar examples [8]. In several papers, it is described how in the case of diseases like Cancer, Mental health and Cardiovascular conditions, scientists are applying predictive algorithms with satisfactory accuracy [9]. An abundance of Machine Learning models exists that can diagnose if a person has Diabetes or is prone to develop Diabetes [10]. However, models that can predict the onset of Diabetes-induced health complications are rarer. One such model used a data mining pipeline to predict T2DM related complications using Electronic Health Record data [11]. Dagliati et al. (2018) predicted complications such as Neuropathy, Nephropathy, and Retinopathy with an accuracy of up to 0.83 [11]. The researchers used few features to conduct the research which reduced the complexity of the model. Furthermore, the model did not predict the complication of Cardiovascular diseases. Recently held studies present data-driven approaches to predict diabetic complications using feature selection [12, 13]. Tanaka et al. (2013) used a statistical model called the Cox regression model to predict the risk of various diabetic complications including Cardiovascular diseases [14]. Nonetheless, apart from these researches, there isn't any significant work that has previously been done regarding the prediction of Diabetes-induced Cardiovascular diseases especially using Machine Learning. One of the likely reasons might be the lack of information and proper variables since heart complications are related to a lot of other factors and variables of the human body [15]. Another major problem is the presence of imbalanced data in different domains of machine learning and data mining especially in medical

science [1, 16]. Mazurowski (2008) presented a paper showing that the classifier performance may deteriorate even with a modest class imbalance in the training data [17]. Even in the field of Diabetes Complication prediction, class imbalance is a concerning issue [11]. To solve the class imbalance problem, data level and algorithm level approaches are usually taken [18]. Guo et al. (2004) proposed a method called The DataBoost-IM approach; it integrates data generation and boosting to overcome class imbalance[19].Learning from imbalanced data has been a focus of intense research and continuous development for more than two decades. With the expansion of machine learning and data mining, more in-depth insight into the nature of imbalanced learning has been gained, adding to the new emerging challenges as well [20]. So even though considering all of these variables can increase the complexity of the model, doing so is very difficult. Regardless, in this paper, these limitations were addressed and focus was put on the imbalance of the dataset in the prediction of risk of Cardiovascular disease alongside Nephropathy.

3 Methodology of Work

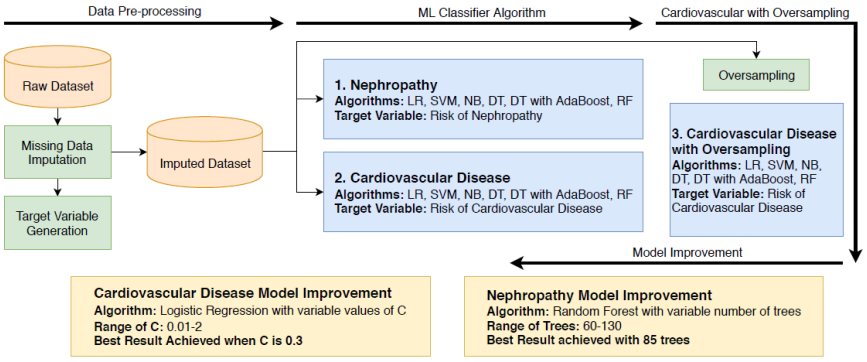


Fig. 1: Methodology of Work

The work presented in this paper can be broadly divided into four sections; i) Dataset Collection, ii) Data Pre-processing, iii) Training and Testing Model and iv) Improving the best ML classifier. In the dataset collection section, an open source dataset was used. Data imputation, Feature Scaling, and Categorical Variable Conversion were done in Data pre-processing. A set of six algorithms were then implemented to find the best possible outcome. After obtaining the best classifier for Nephropathy and Cardiovascular disease, they were further improved by changing a few parameters. Additionally, in the case of Cardiovascular disease, oversampling was also implemented in a particular set of experiments discussed later in the paper. Figure 1 represents the methodology of the work presented. LR, SVM, NB, DT, and RF stands for Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Tree, and Random Forest respectively.

3.1 Dataset

The dataset used in the model is from an open-label, central registration, multi-center, prospective observational study that was conducted at the Tokyo Women's Medical University Hospital in collaboration with 69 other institutions in Japan [21]. It was retrieved from a loyalty free dataset sharing platform and consisted of 779 instances and 164 variables. Out of the 164 variables, 24 were categorical variables, and the rest were numerical variables. Many features are time series data of 1 year at several time differences.

Target Variable Although the dataset contains many features, no feature could be used to determine if a patient is actually at risk of developing either Nephropathy or Cardiovascular disease. Additionally, target variables are needed to measure model performance. Hence, target variables, Risk of Nephropathy and Risk of Cardiovascular Disease have been synthesized. Methods used by former medical researchers have been reapplied for both variables. For the case of Risk of Nephropathy, the conditions (i) Urinary albumin/creatinine ratio greater than 30, (ii) No history of previous renal complication and (iii) GFR less than 60 mL.min-1 have been considered and decisive in all of them has been labeled as '1' [22, 23]. On the other hand, for Risk of Cardiovascular Disease, (i) History of Diabetes Mellitus, (ii) History of Hypertension, (iii) Hypertriglyceridemia and (iv) History of Dyslipidemia have been considered and positive in all of them has been labeled as '1' [24].

Data Imbalance One of the highlighted features of the dataset mentioned above is that it does not have a uniform positive and negative class distribution. For Nephropathy, approximately 25% patients of the dataset were identified as Kidney patients which is an imbalance ratio of 1:4. However, in the case of Cardiovascular disease, the patients having the disease were only about 10% of the set which is an imbalance ratio of 1:10. According to Wong (2009), a rate as low as 1: 10 can be tough to deal with and is inadequate for building a suitable model in most cases [1]. To overcome the issue of imbalance, two different approaches are usually used: Data-level approach and Algorithm-level approach. For the data-level approach, oversampling was applied to the dataset for Cardiovascular Disease prediction, and the performance of the algorithms was compared for balanced and imbalanced dataset. For algorithm-level approach, several classifiers were used including ensemble methods like AdaBoost and Random Forest. The parameters of these classifiers were also tweaked to improve the algorithm and find out the best combination for the imbalanced dataset. Oversampling was implemented by applying the SMOTE algorithm on the training set to make the class ratio 50:50 [26].

3.2 Data Pre-processing

Data Imputation The dataset had values missing at random (MAR), so the missing value of a variable can be predicted from the other variables making

it suitable for imputation. Overall, the maximum number of missing columns for a particular instance was 145 and the minimum was 0. Even though traditional methods of imputation like replacing with mean, replacing with 0 and deleting entire instances were considered, they were found to be inadequate. Instead, imputation was done using an algorithm called missForest with 100 trees [25]. The model yields an out-of-bag (OOB) imputation error estimate which consists of the NRMSE (Normalized Root Mean Squared Error) for the continuous numeric variables and PFC (Proportion of Falsely Classified) for categorical variables. NRMSE, in this case, was 17.89% and PFC was 12.19%.

Feature Scaling and Categorical Variable Conversion Standardization was used as the feature scaling technique as some algorithms do not perform well on unscaled data since variables with higher and lower scale are treated differently. Another essential part of the model is the conversion of the categorical variables to their numerical counterparts to avoid misinterpretation of information from the data. It is implemented by creating dummy variables for each class present in every categorical variable. To prevent the occurrence of a Dummy Variable Trap, n-1 dummy variables were created for a categorical variable with n different values.

3.3 Algorithms and Evaluation Metrics

In this paper, several classification algorithms were used for comparison, to find the best one for each problem. The linear algorithms Logistic Regression, Support Vectors Machines, and Naïve Bayes along with non-linear classifiers Decision Tree and Random Forest were implemented. Additionally, AdaBoost was used for boosting. The primary parameter for logistic regression was L2 regularization with the inverse of regularization strength(C) value being 1. On the other hand for SVM, the RBF kernel was used with the penalty parameter C of the error term being 1. Gaussian Naïve Bayes is the specific algorithm that was applied in the case of Naïve Bayes. CART was the Decision Tree and was utilized with the split criterion based on Gini impurity as one of the non-linear classifiers. Furthermore, AdaBoost was employed using the SAMME.R real boosting algorithm with the maximum number of estimators being 50. Finally, the Random Forest Classifier had 80 trees as the primary number of estimators. The data is imbalanced for both cases, with a prevalence (π) of the disease being 0.249 for Nephropathy and 0.096 for cardiovascular disease. Hence, accuracy is a poor evaluation metric in this case [27]. For the performance evaluation of classifiers, AUC score, Average Precision (AP) and F1 score are usually better metrics. Since it can sort models by overall performance, the AUC is considered more in model assessments. However, the AUC score masks poor performance if the dataset is imbalanced [28]. This is especially the case when the value of π is less than 0.1 [29]. Yuan et al. (2015) have shown this by demonstrating a drastic difference between the AUC and AP score when π changes from 0.5 to 0.1 and less [29]. Therefore, while AUC is a good metric for Nephropathy, it does not work well in case of Cardiovascular Disease. F1 score takes both precision and

recalls into account by taking their harmonic mean with a high score indicating that the model performs better on the positive class [30]. F1 score was prioritized in this paper to evaluate the performance of the algorithms with specific considerations to the precision and recall also. In some medical articles, Average Precision (AP) is considered to assess results.

3.4 Training Dataset

In this paper, cross-validation was done by splitting the overall dataset into training and test set with a ratio of 70:30. For Nephropathy, the test set consisted of 216 instances, 162 belonging to the false class (does not have Kidney complications) and 54 belonging to the true class (has Kidney complications). In the case of Cardiovascular disease, the test set consisted of 234 instances with 211 belonging to the false class (does not have Cardiovascular complications) and 23 belonging to the true class (has Cardiovascular complications).

4 Results

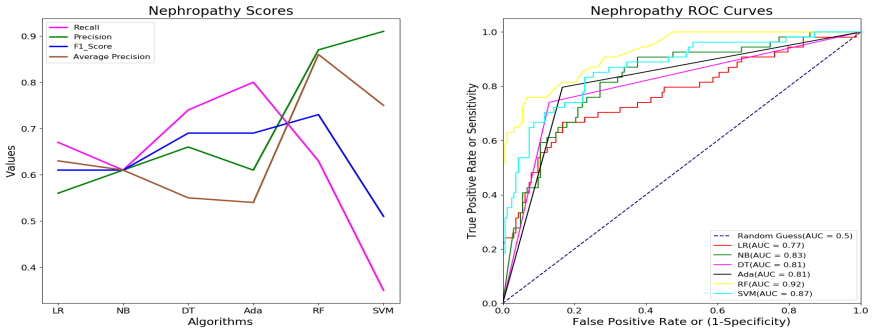
The linear classifiers used were Logistic Regression (LR), Support Vector Machines (SVM) and Naïve Bayes (NB) and the non-linear classifiers used were Decision Tree and Random Forest were implemented. AdaBoost was used since it is capable of handling the class imbalance problem. Oversampling using SMOTE was applied to Cardiovascular Disease. This paper discusses all the models experimented in three subsections: i. Nephropathy, ii. Cardiovascular without Oversampling and iii. Cardiovascular with Oversampling. Each results table contain Accuracy, Precision, Recall, Average Precision (AP), F1 score and AUC score in regard to all algorithms that are being applied.

4.1 Nephropathy

Table 1 shows scores of different performance metrics for all the algorithms for Nephropathy. Figure 2(a) is the graphical representation of the values of different algorithms' recall, precision, AP and F1 score. The difference between a specific metric for each algorithm can be signified. Additionally, Figure 2(b) represents the AUC scores for each algorithm, showing the best and the worst algorithms based on AUC score.

Table 1: Nephropathy Scores

	Logistic Regression (LR)	Support Vector Machines(SVM)	Naïve Bayes (NB)	Decision Tree (DT)	Decision Tree with AdaBoost(Ada)	Random Forest (RF)
Accuracy	0.79	0.83	0.81	0.84	0.82	0.88
Precision	0.56	0.91	0.61	0.66	0.61	0.87
Recall	0.67	0.35	0.61	0.74	0.80	0.63
AP	0.63	0.75	0.61	0.55	0.54	0.86
F1 Score	0.61	0.51	0.61	0.69	0.69	0.73
AUC	0.77	0.87	0.83	0.81	0.82	0.92



(a) Performance Metrics **(b)** ROC curves and AUC scores

Fig. 2: Prediction of Nephropathy (Kidney Disease)

Linear Classifiers: Even though LR and SVM are both linear classifiers, a significant difference can be observed in the performance metrics. The precision is higher in SVM, while the recall and F1 score are better for LR. So while SVM has a higher AUC score of 0.87, it is disregarded. Hence, between these two algorithms, for the prediction of Nephropathy, Logistic Regression is a better option with a recall of 0.67 and F1 Score 0.61 since it classifies the positive class better. Naïve Bayes has a result with the same precision and recall, 0.61. Hence, the AP and F1 score are also 0.61. Further, it can be noticed that the AUC score of NB is 0.83, which is also quite high. However, compared to LR and SVM, the performance of NB cannot be said to be significantly better as none of the results are better than the former algorithms.

Non-linear Classifiers: Decision Tree was used in two ways; with and without the boosting. AdaBoost was used as the booster. The scores in both these cases are similar, with the accuracy, precision, and AP falling slightly for AdaBoost. However, the recall increases to 0.80 for boosted DT while F1 score remains the same at 0.69. Comparing to the linear classifiers, DT has a better overall performance. On the other hand, among all the algorithms, Random Forest has the best AP (0.86), F1 score (0.73) and AUC score (0.92). For disease prediction, recall is an important metric. In this case, the recall is not so high, which is 0.63. However, the accuracy and precision scores of RF is also above 0.85.

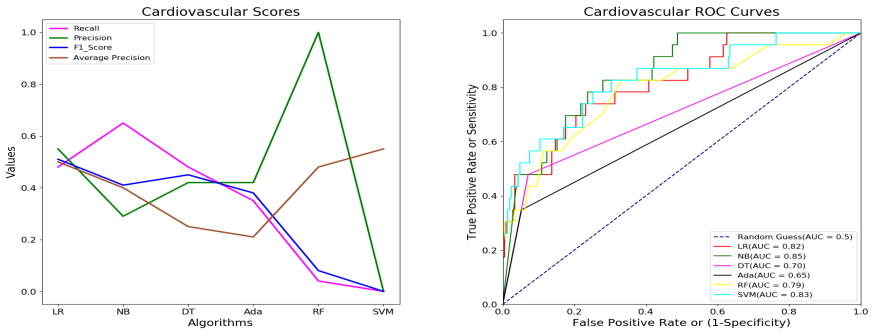
For Nephropathy the best result was given by Random Forest(RF) when taking into account all the performance metrics. Oversampling was not considered for Nephropathy since the data was only slightly imbalanced with 25% belonging to the positive class. RF counters this imbalance by aggregating several decision trees together. Hence, satisfactory results were obtained without the introduction of oversampling. Finally, it can be observed that for Nephropathy, non-linear classifiers work better than linear classifiers as DT, DT AdaBoost and RF gives better F1 Score than LR, SVM, and NB. Since the classes are divided by a non-linear boundary, non-linear classifiers perform better.

4.2 Cardiovascular without Oversampling

Table 2 shows scores of different performance metrics for all the algorithms for Cardiovascular Disease. Figure 3(a) represents the values of different algorithms' performance metrics, and Figure 3(b) represents the AUC scores for each algorithm.

Table 2: Cardiovascular Disease Scores (Without Oversampling)

	Logistic Regression (LR)	Support Vector Machines(SVM)	Naïve Bayes (NB)	Decision Tree (DT)	Decision Tree with AdaBoost(Ada)	Random Forest (RF)
Accuracy	0.91	0.90	0.81	0.89	0.89	0.91
Precision	0.55	NaN	0.29	0.42	0.42	1.00
Recall	0.48	0.00	0.65	0.48	0.35	0.04
AP	0.50	0.55	0.40	0.25	0.21	0.48
F1 Score	0.51	0.00	0.41	0.45	0.38	0.08
AUC	0.82	0.83	0.84	0.70	0.65	0.79



(a) Performance Metrics **(b)** ROC curves and AUC scores

Fig. 3: Prediction of Cardiovascular(Heart) Disease

Linear Classifiers: In the prediction of Cardiovascular disease, LR has a high accuracy of 0.91 and AUC score of 0.82, while having average values for precision (0.55), AP (0.5), recall (0.48) and F1 score (0.51). The results of SVM though are much unexpected as the accuracy and AUC score are high. While on the other hand, the recall is zero and the precision is undefined. Usually, SVM works well for moderate imbalance and performance decreases when it goes towards high imbalance, but this concept can vary with the nature of the dataset [31, 1]. However, for this particular dataset, even though it is moderately imbalanced, the F1 score is zero which means SVM completely fails to predict the positive class. The aforementioned values of recall and precision correspond to only one particular threshold. For other thresholds values, however, the precision and recall are defined giving SVM an average precision of 0.55. Comparing, in this case, LR is the better algorithm. Compared to LR, Naïve Bayes has a higher recall of 0.65 but a lower F1 score of 0.41. Moreover, all the other linear evaluators

have a lower value of F1 score than that of LR. Hence, among all the linear classifiers, Logistic Regression gives the best prediction model.

Non-linear Classifiers: Similar to Nephropathy, DT was implemented in two ways; with and without AdaBoost. The scores in both these cases are similar, though a fall of recall, AP, F1 score and AUC score can be noticed. There is no increase in any of the metrics, and the accuracy and precision are the same for both cases being 0.89 and 0.42 respectively. Comparing to LR, unlike in Nephropathy, all the evaluators have a poorer score, with a steep decrease in AP and F1 score. In the case of RF, it can be noticed that the accuracy is very high (0.91). However, the recall and F1 score are very low with only 0.04 and 0.08 respectively. Therefore, even though the precision is 1, it is a terrible predictor for this case.

However, even though LR is the best predictor for Cardiovascular disease, the precision, recall, AP and F1 score are all very low being close to 0.5 in all cases. The recall is of particular interest in this paper since all the patients who have a chance of developing complications need to be predicted correctly. The recall in most cases are poor since the prevalence of heart patients is less in the dataset. Further, it is seen that the AUC score is not a suitable metric when the prevalence of positive class is extremely low [29]. So in the comparison of the classifiers for this case, more importance was given to the AP score than AUC score. To overcome these issues, oversampling has been used, and the results are discussed in the next subsection.

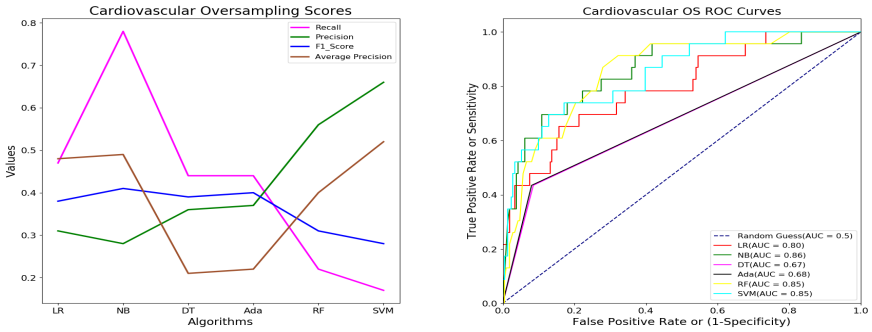
4.3 Cardiovascular with Oversampling

Table 3 shows scores of different performance metrics for all the algorithms for Cardiovascular Disease. Figure 4(a) represents the values of different algorithms' performance metrics, and Figure 4(b) represents the AUC scores for each algorithm.

Table 3: Cardiovascular Disease Scores (With Oversampling)

	Logistic Regression (LR)	Support Vector Machines(SVM)	Naïve Bayes (NB)	Decision Tree (DT)	Decision Tree with AdaBoost(Ada)	Random Forest (RF)
Accuracy	0.58	0.91	0.78	0.87	0.87	0.91
Precision	0.31	0.66	0.28	0.36	0.37	0.56
Recall	0.47	0.17	0.78	0.44	0.44	0.22
AP	0.48	0.52	0.49	0.21	0.22	0.40
F1 Score	0.38	0.28	0.41	0.39	0.40	0.31
AUC	0.80	0.85	0.86	0.68	0.68	0.85

Linear Classifiers: Oversampling the data leads to a decrease in performance for LR shown by a decrease in value of all the metrics. However, in the case of SVM, there is an increase in the value of precision and recall which increases from zero, and hence the F1 score also increases. Since oversampling leads to the prevalence



(a) Performance Metrics **(b)** ROC curves and AUC scores

Fig. 4: Prediction of Cardiovascular (Heart) Disease with Oversampled Dataset

being close to 0.5, the AUC can be considered to measure performance. However, even though SVM has a better accuracy and AUC score than LR, its recall is still lower. The precision is higher for SVM, but a low recall leads to an F1 score of 0.28 which is lower than LR's F1 score of 0.38. Hence, even in this case, LR is a better algorithm than SVM. As mentioned before, the performance of SVM degrades as class imbalance increases. Hence when oversampling is applied SVM performs better as class imbalance problem is resolved. For NB oversampling gives better performance since AP score increases while keeping the F1 score the same. The main observation is that the recall increases significantly to 0.78 which is highest across all combinations for Cardiovascular disease. Since NB has a better AUC and F1 score than LR, it is considered to be better of the two and hence the best linear classifier after oversampling.

Non-linear Classifiers: Just like without oversampling, in oversampling the scores for both DT and Boosted DT are similar. Though Boosted DT has an F1 score which is 0.01 greater than DT, making it slightly better. Overall, oversampling improves the performance of Boosted DT with F1 score increasing from 0.38 to 0.40. For DT, the opposite happens with the performance deteriorating when oversampled. The F1 score decreases from 0.45 to 0.39. Hence it can be deduced that oversampling works for AdaBoosted DT but not DT. The AP score is still poor for both, and since AP is a better measure for the Cardiovascular case, it can be deduced that both classifiers fail. This is due to the linear characteristic of the data for Cardiovascular Disease. Again for RF with oversampling, the accuracy and AUC are very high with both being 0.85 and 0.91 respectively. The recall (0.22) and F1 score (0.31), however, are very low. Even though oversampling improves the values from 0.04 and 0.08, the overall performance is still the same. Therefore, after taking all the performance scores into account, Naive Bayes is the best classifier for the prediction of Heart diseases after oversampling.

To sum up, after comparing both with and without oversampling, Logistic Regression can be considered the best algorithm for classification of Cardiovascular

Disease. Despite Naive Bayes with oversampling having better recall scores, NB's precision is lower, leading to an F1 score of 0.41 which is much lower than LR's F1 score of 0.51. For Cardiovascular Disease, the decision boundary between classes is much more linear with the classes being less well-separated. It leads to poor results of non-linear classifiers due to overfitting of the data.

4.4 Improving Prediction Model

Nephropathy For Nephropathy the best performance in terms of F1 score and recall was given by Random Forest with 80 trees. It is possible to obtain the right balance between performance, processing time, and memory usage when the number of trees in a forest is between 64 and 128 [32]. The Random Forest algorithm was implemented on a variable number of trees within the range of 60 to 130 to find which one maximizes the performance. It can be seen from Figure 5 that best performance is given by 85 trees with the F1 score, recall, precision, and AP being 0.75, 0.65, 0.89 and 0.86 respectively. In general, it can be observed from Figure 5 that the value of the metrics' fluctuates up and down until it reaches 100 trees and then for all parameters, the values stabilizes.

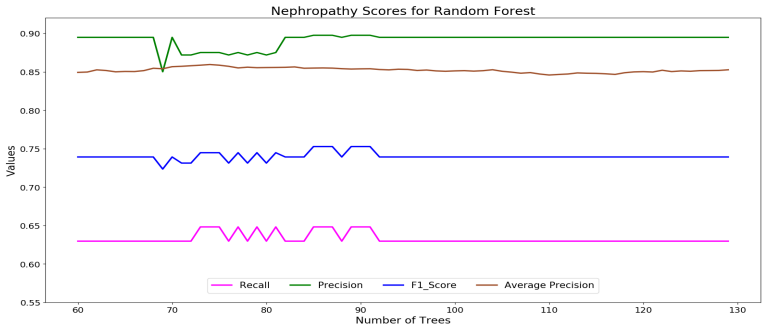


Fig. 5: Improved Nephropathy scores for Random Forest

Cardiovascular Disease In the case of Cardiovascular Disease, the best performance in terms of both F1 score and recall was given by Logistic Regression without oversampling. The inverse of regularization strength(C) in this case was 1. As regularization strength increases, the model generalizes better by taking slightly useful features into account. Since C is the inverse of the regularization strength, as C decreases regularization increases. Hence by changing the value of C, the performance of the Logistic Regression model can be improved. To test this, Logistic Regression algorithm was applied with C values ranging from 0.01 to 2. From Figure 6 it can be seen that the best performance when considering all metrics comes when the value of C is 0.3. At this point, the F1 score, recall, precision, and AP are 0.54, 0.48, 0.61 and 0.53 respectively. Even though recall remains the same, the value of the other metrics' increases. The most significant change is in precision which increases by 0.06. Overall, it can be seen from Fig.

8 that with a C value close to 0, the precision is very high and the recall is very low, leading to a poor F1 score. However, as the C value is increased the F1 score increases, maximizing at around 0.3, which is where both recall and precision Scores are acceptable. Increasing C even further leads to the value of all the metrics either stabilizing or decreasing even more.

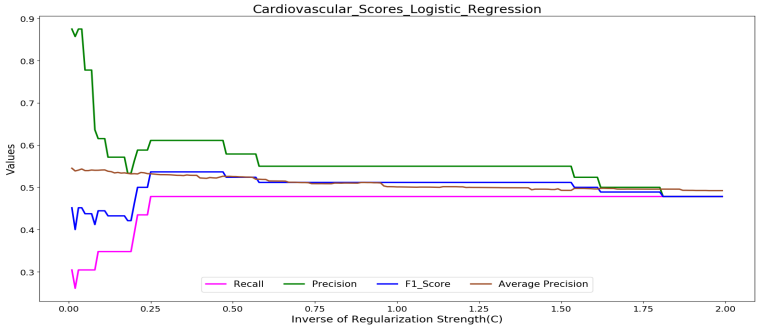


Fig. 6: Improved Cardiovascular Disease scores for Logistic Regression

5 Discussion

In this paper, the best classifier for Nephropathy in terms of F1 score and AUC is Random Forest with 85 trees. Furthermore, in the case of Cardiovascular Disease, Logistic Regression is the best classifier without oversampling in terms of F1 Score and AUC. Among all the classifiers applied, SVM showed high precision and Decision tree had the best recall score for predicting the risk of Nephropathy. However, Random Forest provided the highest AUC and F1 Score, so it is more logical to select RF as the best classifier for prediction of Nephropathy onset. For Nephropathy, the dataset had an imbalance of ratio 1:4. From the results, it is evident that this data imbalance did not degrade the performance of the classifiers. Moreover, the best result was given by Random Forest, an ensemble method which is better at handling imbalanced data since it runs several Decision Tree classifiers and aggregates the result. On the other hand, in the case of Cardiovascular Disease prediction the imbalance ratio was 1:10, hence oversampling was used. The best result was obtained without applying Oversampling: Logistic Regression with the highest F1 score. Again the Recall score was not very high owing to less prevalence of patients in the dataset. Cardiovascular Disease onset prediction was further explored with the addition of oversampling. Nevertheless, oversampling failed to provide much improvement to the result. This was due to overfitting the data when the minority class was oversampled. Also, the optimal class distribution is not known; in this paper, it was assumed to be 50:50 which may not be the case. SVM showed deficient F1 score of only 0.28, but the precision was higher than that of Logistic Regression without oversampling. Even though SVM had higher AUC score after oversampling, it was disregarded since the AUC score can be misleading in cases of low

prevalence. So, although AUC is a useful metric for Nephropathy, it does not work well in case of Cardiovascular Disease. It is noticed that nonlinear classifiers perform better while predicting Nephropathy onset in comparison with Cardiovascular Disease prediction. However, it was found that since 1:4 and 1:10 are not extreme imbalance cases, the results are better when imbalance is ignored. Furthermore, in the later portion of the paper, Random Forest (for Nephropathy) and Logistic Regression without Oversampling (for Cardiovascular Disease) were further tuned to improve the results more. In the case of Random Forest, a different number of estimators within the range of 60 to 130 trees was applied to find which one maximizes the performance. It has been observed that 85 trees maximize Recall, Precision, and F1 Score. On the other hand, for Cardiovascular disease prediction, the initial value of the inverse of regularization strength(C) was 1. C values ranging from 0.01 to 2 was exercised and the best results are obtained when the value is 0.3. Value of both F1 score and Recall increases and the highest increment is observed in precision which increases by 0.06. It can be seen from Figure 5 and Figure 6 that the performance metrics become horizontal lines after a specific value of C (for Logistic Regression) and a particular number of trees (for Random Forest). A couple of interesting observations can be made from this research. Firstly, the model worked better for Nephropathy than it did for Cardiovascular Disease, even though the same dataset is used for both; implying that the variables in the dataset are more inclined towards Nephropathy than Cardiovascular Disease. Further, the prevalence of the disease also played a significant part in the results. Since Nephropathy had 25% positive cases compared to a mere 10% for Cardiovascular Disease, the results were better for Nephropathy. Secondly, although the dataset had 164 features, it had only 779 instances, the model gave encouraging results.

6 Conclusion

This paper explains how Machine Learning based linear and non-linear classifiers can be adopted in clinical diagnostics to create systems that use patient-specific information to predict the probability of Diabetes-induced complications. A total of five classifiers have been applied here on an imbalanced dataset and the results have been measured using various performance metrics. Random Forest and Logistic Regression provided the best results for the prediction of Nephropathy and Cardiovascular Diseases respectively. Furthermore, Random Forest has been applied multiple times, each time with a different number of estimators and the classifier works best with 85 trees. Similarly, Logistic Regression delivers the best result when the inverse of regularization strength, C, is 0.3 amongst all separate values that were tried. This research work was quite challenging owing to various facts. A significant obstruction was the missing values present in the current dataset. Although the missing data were imputed using a widely used data imputation algorithm, nevertheless, accuracy of all the imputed missing values may not be entirely correct, according to the medical aspect. Regardless of that, the model can satisfactorily predict the onset of Diabetes-induced Nephropathy and Cardiovascular Disease. This paper also reflects how the clinical data,

which are usually imbalanced, affects the prediction system. It is certain that in the foreseeable future a predictive model like this can be successfully used for prognosis, diagnosis and treatment planning of ever-increasing Diabetic patients within clinical information systems.

References

1. Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.
2. Mani, S., Chen, Y., Elasy, T., Clayton, W., and Denny, J. (2012). Type 2 diabetes risk forecasting from EMR data using machine learning. In AMIA annual symposium proceedings (Vol. 2012, p. 606). American Medical Informatics Association.
3. Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., and Malanda, B. (2018). IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice, 138, 271-281.
4. Fowler, M. J. (2008). Microvascular and macrovascular complications of diabetes. Clinical diabetes, 26(2), 77-82.
5. Gregg, E. W., Li, Y., Wang, J., Rios Burrows, N., Ali, M. K., Rolka, D., et al. (2014). Changes in Diabetes-related complications in the United States, 1990–2010. New England Journal of Medicine, 370(16), 1514-1523.
6. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine, 24(9), 1342.
7. AbuKhousa, E., and Campbell, P. (2012, March). Predictive data mining to support clinical decisions: An overview of heart disease prediction systems. In Innovations in information technology (iit), 2012 international conference on (pp. 267-272). IEEE.
8. Yun, 15 July 2018, Chinese AI Beats Doctors in Diagnosing Brain Tumors.
9. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine Learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.
10. Barakat, N., Bradley, A. P., and Barakat, M. N. H. (2010). Intelligible Support Vector Machines for diagnosis of Diabetes mellitus. IEEE transactions on information technology in biomedicine, 14(4), 1114-1120.
11. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., et al. (2018). Machine Learning methods to predict Diabetes complications. Journal of Diabetes science and technology, 12(2), 295-302.
12. Liu, B., Li, Y., Sun, Z., Ghosh, S., and Ng, K. (2018). Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-task Survival Analysis Approach.
13. Cho, B. H., Yu, H., Kim, K. W., Kim, T. H., Kim, I. Y., and Kim, S. I. (2008). Application of irregular and unbalanced data to predict diabetic Nephropathy using visualization and feature selection methods. Artificial intelligence in medicine, 42(1), 37-53.
14. Tanaka, S., Tanaka, S., Iimuro, S., Yamashita, H., Katayama, S., Akanuma, Y., et al. (2013). Predicting macro-and microvascular complications in type 2 diabetes: the Japan Diabetes Complications Study/the Japanese Elderly Diabetes Intervention Trial risk engine. Diabetes Care, DC_120958.

15. Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837-1847.
16. Rahman, M. M., and Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing, 3(2), 224.
17. Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, 21(2-3), 427-436.
18. Ali, A., Shamsuddin, S. M., and Ralescu, A. L. (2015). Classification with class imbalance problem: a review. Int J Adv Soft Comput Appl, 7(3), 176-204.
19. Hongyu Guo, Herna L. Viktor: Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. SIGKDD Explorations 6(1): 30-39 (2004)
20. Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221-232.
21. Tomonaga, O. (2017, April 27). JAMP_DATA0722figshaer.xlsx (Version 1). [Retrieved from: doi.org/10.6084/m9.figshare.4924037.v1].
22. Gross, J. L., De Azevedo, M. J., Silveiro, S. P., Canani, L. H., Caramori, M. L., and Zelmanovitz, T. (2005). Diabetic Nephropathy: diagnosis, prevention, and treatment. Diabetes care, 28(1), 164-176.
23. Mogensen, C. E. (1987). Microalbuminuria as a predictor of clinical diabetic Nephropathy. Kidney international, 31(2), 673-689.
24. Han, S. H., Nicholls, S. J., Sakuma, I., Zhao, D., Koh, K. K. (2016). Hypertriglyceridemia and Cardiovascular Diseases: Revisited. Korean Circ J, 46(2), 135–144. Doi: 10.4070/kcj.2016.46.2.135.
25. Stekhoven, D. J., and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
26. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
27. He, H., and Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, (9), 1263-1284.
28. Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013, September). Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on (pp. 245-251). IEEE.
29. Yuan, Y., Su, W., and Zhu, M. (2015). Threshold-free measures for assessing the performance of medical screening tests. Frontiers in public health, 3, 57.
30. Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced datasets. Journal Of Information Engineering and Applications, 3(10).
31. Wu, G., and Chang, E. Y. (2003, August). Class-boundary alignment for imbalanced dataset learning. In ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC (pp. 49-56).
32. Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012, July). How many trees in a random forest?. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 154-168). Springer, Berlin, Heidelberg.