An enhanced predictive modelling framework for highly accurate non-alcoholic fatty liver disease forecasting

Nidhi Arora[1], Shilpa Srivastava[2], Aprna Tripathi[3], Varuna Gupta[2]
[1]Department of Computer Science, Kalindi College, University of Delhi, Delhi, India
[2]School of Sciences, Christ University, Bengaluru, India
[3]Department of Data Science and Engineering, Manipal University Jaipur, Jaipur, India

Article Info

Article history:

Received Jun 14, 2024
Revised Aug 26, 2024
Accepted Aug 31, 2024

Keywords:

Chronic diseases
Class imbalance
Classification
NAFLD
Predictive modelling
SMOTE

ABSTRACT

Non-alcoholic fatty liver disease (NAFLD) is a chronic medical ailment characterized by accumulation of excessive fat in the liver of non-alcoholic patients. In absence of any early visible indications, application of machine learning based predictive techniques for early prediction of NAFLD are quite beneficial. The objective of this paper is to present a complete framework for guided development of varied predictive machine learning models and predict NAFLD disease with high accuracy. The framework employs step–by-step data quality enhancement to medical data such as cleaning, normalization, data upscaling using SMOTE (for handling class imbalances) and correlation analysis-based feature selection to predict NAFLD with high accuracy using only clinically recorded identifiers. Comprehensive comparative analysis of prediction results of seven machine learning predictive models is done using unprocessed as well as quality enhanced data. As per the observed results, XGBoost, random forest and neural network machine learning models reported significantly higher accuracies with improved ‘AUC’ and ‘ROC’ values using preprocessed data in contrast to unprocessed data. The prediction results are also assessed on various quality metrics such as ‘accuracy’, ‘f1-score’, ‘precision’, and ‘recall’ significantly support the need for presented methodologies for qualitative NAFLD prediction modelling.

Corresponding Author:

Nidhi Arora
Department of Computer Science, Kalindi College, University of Delhi
Delhi 110008, India
Email: nidhiarora@kalindi.du.ac.in

1. INTRODUCTION

Chronic metabolic disorders are disorders that are long lasting and are caused due to disruptions in the metabolism of human body [1]. These categories of diseases are primarily not indicative using general physical and clinical assessments, but lead to severe impacts on the health in the later stages of life. Often one metabolic disorder lead to other disorders and can have multiorgan impacts. Non-alcoholic fatty liver disease (NAFLD) is one of the complex metabolic chronic disorder that has recently seen a surge. The disease is coupled with other metabolic disorders as well such as type 2 diabetes mellitus and obesity [2]. NAFLD condition in a human being is indicated by the build-up of excessive fat content in the liver of those individuals who are not consuming alcohol significantly. The complexity of NAFLD lies with the influence of multifactorial impacts in the light of multiple genes coming together in respect to many environmental factors. It is difficult to identify and treat such multifactorial diseases as specific factors which lead to such diseases are still under the wrap for medical researchers.

Medical informatics, is an interdisciplinary field of research which is focused on application of varied information technology deriven artificially intelligent methodologies for the diagnosis and treatment of diseases [3], [4]. The field has garnered much attention in the research spectrum of medical as well as computer science professionals. With the rise in the availability of substantial medical information due to information technology advancements, researchers are quite intelligently assessing this information to support decision making for health professionals by application of varied data analytics, data mining and machine learning techniques under the umbrella of artificial intelligence (AI). Of many AI techniques, machine learning and predictive modelling has garnered much attention of researchers of late [5] in this field. Machine learning focus on utilizing available medical data of patients such as medical records, significant test results, and medical prescriptions. for making machines learn significant patterns and develop its own prediction models. Such prediction models can then be utilized for automatic prediction of many diseases quite early. Such techniques and their usefulness can be super beneficial for timely prediction of possibility of metabolic disorders, for which observable health indicators appear quite lately in patient’s life. Machine learning based predictive analysis by training classification and regression models has been done for varied diseases such as diabetes, heart diseases, liver diseases to name a few [6], [7]. However, the training of such predictive models and accuracy of prediction results largely depends on the quality and quantity of data available for training and testing of prediction models.

Disease specific application of chosen machine learning techniques have been tried and evaluated by many researchers in recent times. A dedicated study for prediction of cardiovascular diseases was conducted by Krittanawong *et al.* [8]. The study outlined the importance of support vector machines (SVM) and boosting algorithms in prediction of the concerned diseases with higher AUC metric results. Hybridized machine learning algorithms and their effectiveness in prediction of heart diseases have been done by Mohan *et al.* [9]. The authors presented the results of a new prediction model which was a hybrid random forest with a linear model (HRFLM) with an accuracy of 88.7%. Many other researchers [10], [11] also evaluated prediction models for heart diseases using varied machine learning techniques on different sample medical datasets.

Another category of diseases is neurogenerative, for which some latest studies conducted by [12], [13] have shown utilization of machine learning based predictive methodologies to develop prediction models. Prediction models developed for Rheumatic diseases have been explored by [14], [15]. In recent times, developing prediction models for chronic and metabolic disorders have also garnered much attention. In other research, Nusinovici *et al.* [16] also assessed the functioning of machine learning algorithms for the risk prediction of heart (cardiovascular) diseases, kidney disease, hypertension and diabetes using simple clinical predictors. Many other studies conducted by various researchers for chronic and metabolic diseases have been done recently using varied machine learning classification algorithms.

Another metabolic disease which has garnered attention of researchers in medical informatics is NAFLD disease. NAFLD disease may lead to other metabolic diseases such as cardiovascular diseases [17] and generate many other health risks. AI techniques including machine learning such as classification using logistic regression, random forest, XGBoost and decision tree are applied for detecting this disease by researchers as presented by Wong *et al.* [18]. The applied techniques used datasets of electronically stored health records of individuals, Biopsy records of liver, along with liver images. Various other researchers have also used clinical observant parameters-based dataset for developing machine learning models and predict NAFLD disease [19], [20]. The accuracy results as discussed in above studies are varied for different machine learning models. The results reported are enlightening however it is evident that not much analysis is done regarding the significance of processing of datasets to improve its quality as an important step prior to development of machine learning prediction models for prognosis of NAFLD diseases.

The current research surge in the prediction of metabolic and chronic diseases is clearly visible due to high health impacts of such diseases. However, analyzing available data sets prior to development of predictive models have not been done in depth. Such preprocessing for quality improvements of the available datasets can have a significant impact of training prediction models in broad ranges of machine learning and AI on a whole. Therefore, this paper presents a complete framework for preprocessing for quality improvement of medical datasets and then evaluating these high-quality data sets in development of machine learning trained highly qualitative prediction models. Varied methodologies are utilized for quality improvements and then the improved dataset is employed for training of machine learning models. The paper performs in-depth experimental evaluations of training and testing accuracies of seven prevelant machine learning models. The conducted studies have strengthened the fact that preprocessing of datasets plays a significant role in not only improving accuracies but also improving results of other metrics such as AUC and F1-score.

Research gap and main contribution: availability of quality data sets is quite an essential requirement for training of prediction models. However, in medical ecosystems, data is enormously generated on a regular basis but primarily retained in raw and unprocessed form [21]. More so, a common framework is required to store patient id indexed medical datasets in a common server linking all the medical hospitals throughout as proposed by Arora *et al.* [22]. Data preprocessing, an important ingredient of this framework, has to be assessed for medical data sets and common policies for preprocessing medical data to improve quality for developing highly efficient prediction models is required. This paper explores and presents varied data preprocessing-based methodologies for medical data to predict NAFLD metabolic disorder. The methodologies applied have successfully improved not only the accuracy of the prediction models but also the precision of the trained model. Varied techniques for data quality improvement have been rigorously applied and results presented in this paper which can be used in the medical informatics. Training and testing data prediction results on varied prediction quality quantification metrics such “accuracy”, “precision”, “recall” and “F1-score” using unprocessed data and processed data has been documented for seven different state of the art machine learning based predictive models named “logistic regression”, “Naïve Bayes”, “SVM”, “decision tree”, “neural network”, “random forest” and “KNN”. Varied python-based utilities to automate the quality enhancement process have been listed such as use of python imputer and python smote utility to enhance and balance the data sets in case of imbalance class distribution for quality enhancements and realistic predictions using equal distribution of class data sets.

2. METHOD

Clinical observations though are not directly indicative of NAFLD, but if such observations are recorded overtime with their prediction of NAFLD status of diagnosed patients, then such a data can certainly be utilized as classified data for training machine learning predictive models. However, since the clinical observations are primarily recorded manually and largely depends on medical experts; the data generally suffers from incompleteness; redundancies; unequal class distributions and multiple columns of clinical variables which are highly correlated. Dataset with such properties; may behave inconsistently once statistically utilized for training machine learning predictive analysis. The primary objective of this conducted research work is to examine and utilize data preprocessing techniques prior to application of model training phase as a compulsory step in development of trained prediction models of metabolic and chronic disorders. The methodology for the conducted research works towards generating highly qualitative, clean and evenly distributed data for developing qualitative predictive models, which report high performance levels in terms of accuracy, precession, recall, F1-score, and AUC prediction quality quantifier metrics. The ideology is not only to improve accuracy but also to improve precision values and AUC results, which is a significant indicator of highly accurate prediction. The methodology adapted for the conducted research work is depicted using a framework as shown using phase diagram in Figure 1. The broad steps as shown in above phase diagram are explained in Figure 1.

- Step 1. Data set generation of NAFLD disease: as a first step in the suggested phased diagram, clinical observations from different medical sources are gathered and merged together to form a single repository of a NAFLD disease. For our conducted study, a clinical observations dataset of 17549 individuals merged from different sources available on vincentarelbundock.github.io/Rdatasets/datasets.html is used in the conducted study. For seven attributes recorded (id attribute is irrelevant) and available on platform is used in the conducted study. A snapshot of the dataset is shown in Figure 2.

The primary reference for the data set shown above for NAFLD roots back to the study conducted by Allen *et al.* [23]. The author devised a population cohort consisting of the data collected for all adult NAFLD prospective patients from years spanning in the range of 1997 to 2014. The snapshot shown above is regenerated such that all the columns’ data in the original study are remapped to hide and protect patient’s confidentiality. Patient’s age represented in the Figure 2 is the numeric age at the date of indexing, and the subject identifier is a random number. As a final data protection for individuals, the dataset shown above and utilized in our conducted study uses only approximately 90% randomly picked data from the original data of the authors of the primary study. The description of the variable names used in the data set are as follows: ‘id’ is the subject’s (individual’s) identifier; ‘age’ is the age at date of indexing to the conducted study ; ‘male’ variable stores value ‘0’ for representing females and 1 otherwise; ‘weight’ represents weight of the subject in kilograms; ‘height’ column stores the height of the subject under study in centimeters; ‘bmi’ is the body mass index of the subject; ‘case.id’ is the id number of the NAFLD case with which subject under study gets matched; ‘futime’ is the time to death or last follow-up of the subject; ‘status’ is 0 if the subject was alive at the last follow-up else 1 if the subject was dead.

- Step 2. Missing data amputations: the data snapshot depicted in Figure 2 displays the presence of missing data values in the recorded observations of individuals represented as ‘NAN’ in numerical columns (‘weight’, height’, ‘bmi’). Data having missing values, is not suitable for statistical analysis nor for machine learning predictive modelling. Hence, such values need critical analysis and amputations using suitable methodologies. For our conducted study, missing values in the datasets are mean replaced using a

SimpleImputer utility from sklearn.impute package of python [24]. The corrected dataset snapshot by mean replacement using above python utility for numerical columns 'NAN' values is shown in Figure 3.

- Step 3. Data correlation evaluation: predictive modelling for classification of dependent variable (class variable) highly depends upon number of independent variables used in the model generation. High number of variables used in the prediction model often lead to problem of overfitting which lead to false and over positive predictions on training while having very poor performance on testing dataset. A quite simpler method to handle this problem is to keep minimum independent variables in predictive model generation, by dropping highly correlated variables. Hence, a correlation heat map matrix is generated as shown in Figure 4 for NAFLD dataset, which clearly depicts high correlation between 'weight' and 'bmi'. Hence, 'bmi' is dropped from the dataset to retain only significantly highly uncorrelated features which can have independent impact on model development and reduce the complexity of model.

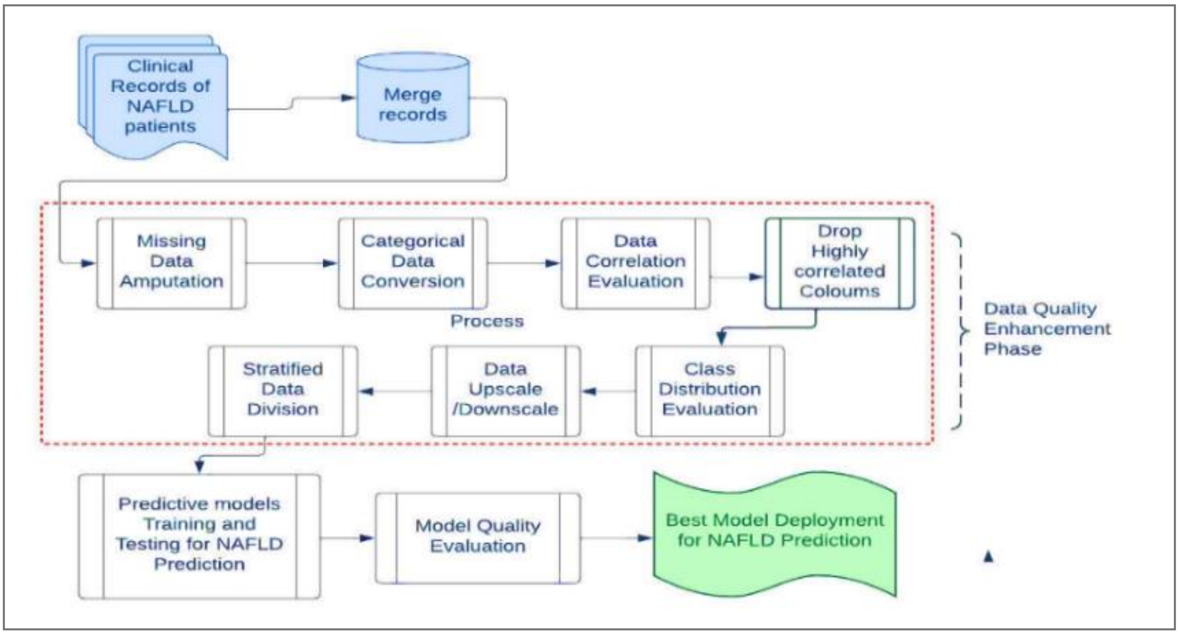
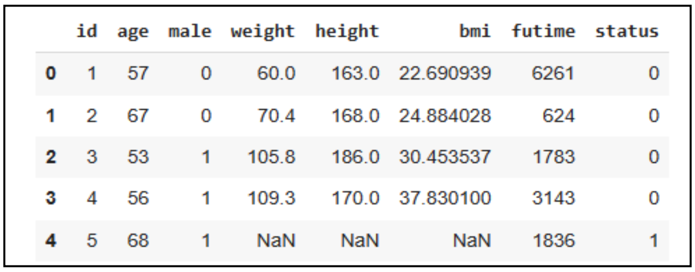
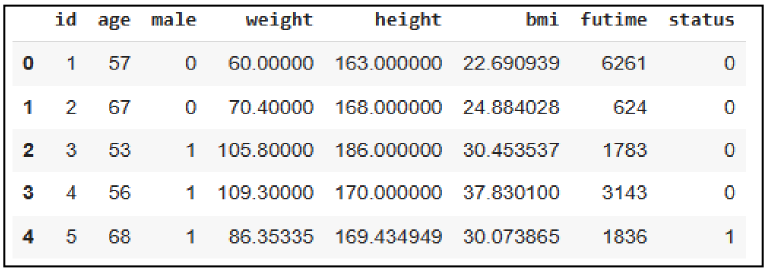

Clinical Records of NAFLD patients
Merge records
Missing Data Amputation
Categorical Data Conversion
Data Correlation Evaluation
Drop Highly correlated Coloums
Process
Data Quality Enhancement Phase
Stratified Data Division
Data Upscale /Downscale
Class Distribution Evaluation
Predictive models Training and Testing for NAFLD Prediction
Model Quality Evaluation
Best Model Deployment for NAFLD Prediction

Figure 1. Phase diagram for NAFLD disease prediction using qualitative data



	id	age	male	weight	height	bmi	futime	status
0	1	57	0	60.0	163.0	22.690939	6261	0
1	2	67	0	70.4	168.0	24.884028	624	0
2	3	53	1	105.8	186.0	30.453537	1783	0
3	4	56	1	109.3	170.0	37.830100	3143	0
4	5	68	1	NaN	NaN	NaN	1836	1

Figure 2. NAFLD dataset snapshot



	id	age	male	weight	height	bmi	futime	status
0	1	57	0	60.00000	163.000000	22.690939	6261	0
1	2	67	0	70.40000	168.000000	24.884028	624	0
2	3	53	1	105.80000	186.000000	30.453537	1783	0
3	4	56	1	109.30000	170.000000	37.830100	3143	0
4	5	68	1	86.35335	169.434949	30.073865	1836	1

Figure 3. Corrected data with mean replacement

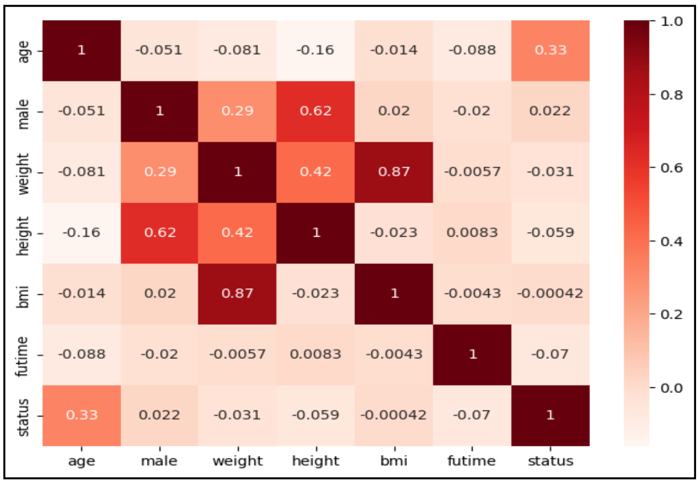
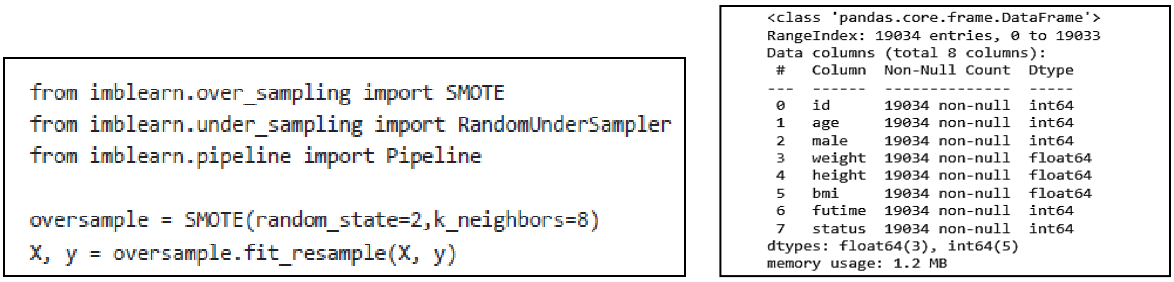


Figure 4. Correlation Heatmap for NAFLD dataset

- Step 4. Class balance distribution evaluation: it is observed in the NAFLD dataset that the class variable ‘status’ has quite unequal distribution for values ‘0’ and ‘1’. It is having more of ‘1’ as compared to ‘0’, which make the predictive modelling biased for class value ‘1’. Hence, to make the training of predictive models unbiased, a necessary step to equalize the class distribution, with performing up/down-sampling is done. In up-sampling, rows for class value with lower occurrence are appended along with majority class rows in the dataset with an overall distribution representing the original dataset. The data can be under-sampled with a reverse procedure. Different utilities in python are available for up-sampling and under-sampling as shown in below Figure 5. Figure 5(a) shows the usage of python SMOTE [25] utility to automatically up-sample the minority class. This utility is utilized in our conducted research work. Figure 5(b) shows the details of dataset having now 19034 total rows with oversample data.



```
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
from imblearn.pipeline import Pipeline

oversample = SMOTE(random_state=2,k_neighbors=8)
X, y = oversample.fit_resample(X, y)
```

(a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19034 entries, 0 to 19033
Data columns (total 8 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      19034 non-null  int64
 1   age     19034 non-null  int64
 2   male    19034 non-null  int64
 3   weight  19034 non-null  float64
 4   height  19034 non-null  float64
 5   bmi     19034 non-null  float64
 6   futime  19034 non-null  int64
 7   status  19034 non-null  int64
dtypes: float64(3), int64(5)
memory usage: 1.2 MB
```

(b)

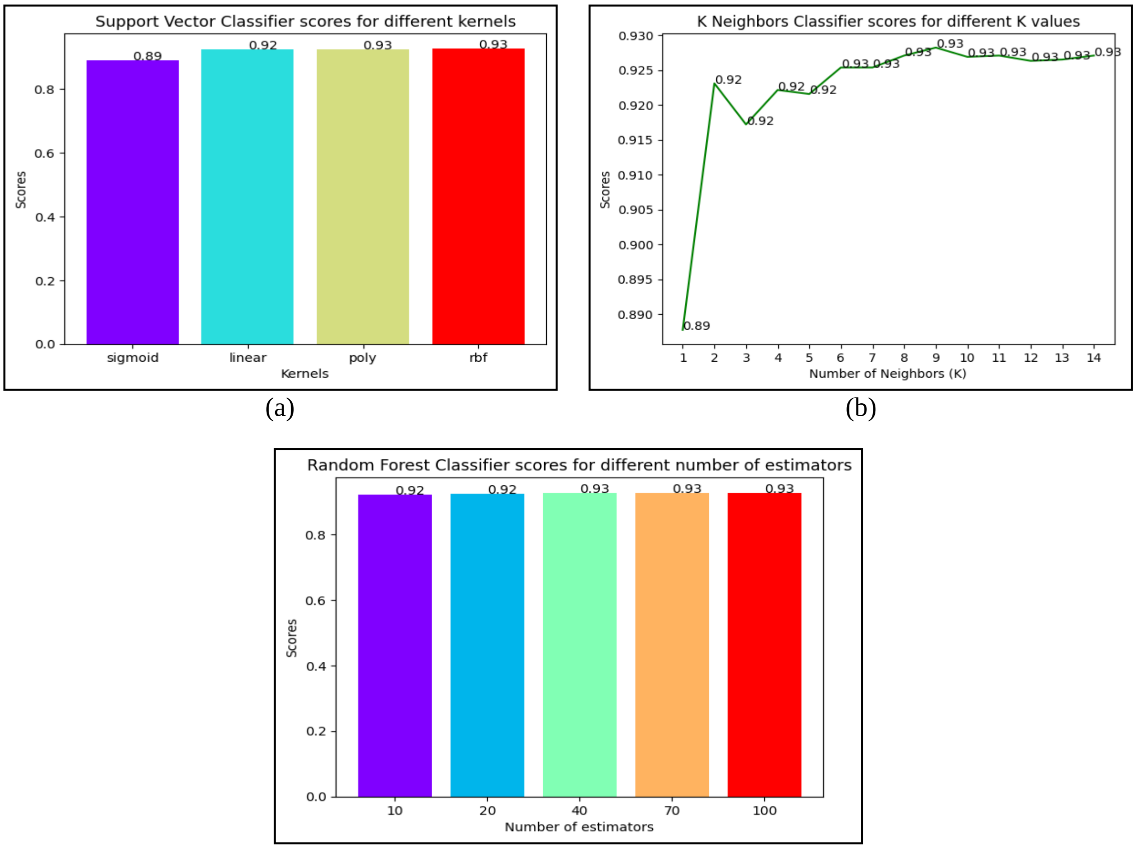
Figure 5. Upsampling of data for handling class imbalance (a) usage of SMOTE utility for oversampling of NAFLD dataset and (b) upscaled NAFLD dataset details

- Step 5. Stratified sampling: stratified sampling is an essential step to divide the data set into training and testing data; so as to have equally distributed representation of all the classes of dependent variable (in our case its ‘status’) in both the training as well as testing dataset. The stratified distribution ensures justified training model development which is unbiased having realistic testing accuracy results.

3. RESULTS AND DISCUSSION

Experimentation using the processed and upscaled data of NAFLD by following the above phased procedure step by step is done using Google Collab. Results of prediction accuracy of the trained model on 30% testing dataset is evaluated using non-stratified, stratified and upscaled stratified data. Confusion matrices are generated to quantify accuracy, precision, recall, F1-score metric values. These metrics are very appropriate indicators of quality of predictions done by any predictor models. Results of prediction of seven predictive models on test data prediction for NAFLD disease are collected and analyzed in our conducted research work. The seven predictive model generation algorithms are logistic regression, Naïve Bayes, SVM, neural network, random forest, KNN, and decision tree.

Parameter tuning: for our conducted experimentation, a significant step to tune the 'kernel' parameter of SVM model, 'number of neighbors' parameter in KNN model and 'number of estimators' parameters in random forest model is also performed. The results of accuracy scores obtained for this step are depicted using various plots in Figure 6. It is observed that results of accuracy were best at Poly kernel for SVM as shown in Figure 6(a), at 8 neighbors for KNN as shown in Figure 6(b) and 40 estimators for random forest models as shown in Figure 6(c). Therefore, for conducting further experimentations, these parameters are chosen and fixed.



(a)

(b)

(c)

Figure 6. Results of different parameter tuning; (a) parameter tuning for kernels in SVM, (b) parameter tuning for number of neighbours parameter in KNN and (c) parameter tuning for number of estimators in random forest

The detailed testing results obtained on non-stratified partition based NAFLD data set are shown in Table 1. The results show best accuracy level of 93 percent for Naïve Bayes model. For most of the models the accuracy was round off 93 percent, except decision tree. The precision values reported were low for decision tree and KNN. Table 2 show the testing results on NAFLD dataset after using stratified sampling. Not much improvement is observed in accuracy values of the prediction, however precision levels are improved for all the predictive models.

Table 1. Prediction quality results of non-stratified quality improved NAFLD test data

Predictive model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.928	0.632	0.93	0.90
Naïve Bayes	0.931	0.681	0.93	0.91
SVM (Poly)	0.925	1.0	0.93	0.89
Neural network	0.928	0.55	0.93	0.91
Decision tree	0.879	0.22	0.88	0.88
KNN (8 neighbors)	0.927	0.596	0.93	0.90
Random forest (40 estimators)	0.925	0.514	0.93	0.91

Table 2. Prediction quality results of stratified quality improved NAFLD test data

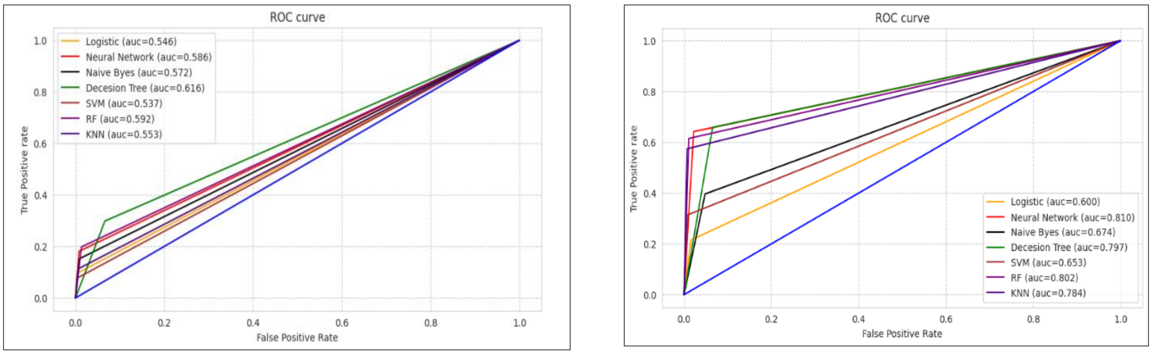
Predictive model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.926	0.69	0.93	0.90
Naïve Bayes	0.926	0.607	0.93	0.91
SVM (Poly)	0.922	1.0	0.92	0.89
Neural network	0.929	0.717	0.93	0.91
Decision tree	0.884	0.275	0.88	0.89
KNN (8 neighbors)	0.925	0.621	0.93	0.90
Random forest (40 estimators)	0.925	0.554	0.93	0.91

Table 3 shows the testing results using stratified distributed and Upscaled NAFLD data set. A quite interesting observation for the results obtained indicate a balanced result for accuracy as well as precision. A balanced result of accuracy and precision lead to better ROC curves with higher AUC values, which is a strong indication for the correctness of the developed predictive model using Upscaled and quality enhanced dataset. ROC curves for stratified data for the results of Table 2 and Table 3 are plotted and shown in Figure 7.

Table 3. Prediction quality results of stratified quality improved upscaled NAFLD test data

Predictive model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.869	0.704	0.87	0.84
Naïve Bayes	0.868	0.591	0.87	0.86
SVM (Poly)	0.889	0.859	0.89	0.87
Neural network	0.927	0.836	0.93	0.92
Decision tree	0.893	0.639	0.89	0.89
KNN (8 neighbors)	0.930	0.937	0.93	0.92
Random forest (40 estimators)	0.933	0.908	0.93	0.93

ROC curves: ROC curves are also generated as shown in Figure 7. Figure 7(a) depicts the plot for stratified data, whereas Figure 7(b) shows the plot for stratified upscaled data. It is evident that ROC curve shown in Figure 7(b) has higher AUC values, therefore the results of accuracy and precision generated for stratified distributed upscaled data are highly reliable and indicate realistic accuracy for the future predictions. The highest AUC is reported for neural network model and random forest model.



(a) (b)

Figure 7. ROC and AUC values for (a) stratified distributed data and (b) stratified distributed upscaled data

4. KEY TAKE AWAYS

Readers will be able to understand the significance of data preprocessing to medical data for accurate prediction in this paper. Training and testing data prediction results on varied prediction quality quantification metrics such "accuracy", "precision", "recall", and "F1-score" using unprocessed data and processed data has been documented for seven different state of the art machine learning based predictive models named "logistic regression", "Naïve Bayes", "SVM", "decision tree", "neural network", "random forest", and "KNN". Readers will also be able to identify experimental results of varied python-based utilities to automate the quality enhancement process. Python utilities presented here are python imputer and python smote utility to enhance and balance the data sets in case of imbalance class distribution for quality enhancements and realistic predictions using equal distribution of class data sets.

5. CONCLUSION

Predictive analysis is a significant area of research in medical informatics specifically for developing machine learning models for chronic diseases. NAFLD is one such disease which is impacting quite a large number of individuals now a days. High quality predictive models having high quality indicators for prediction accuracy. Such models, if developed, will be a boon for all stakeholders in the health sector. However, due to scarcity of high-quality real time clinical datasets, quality prediction with high accuracy is a challenge. This paper has presented a modular approach with systematic phased application of data quality enhancement steps for NAFLD clinical observation data sets using python simple utilities such as imputer and SMOTE. SMOTE utility utilizes generative AI methodologies to upscale the imbalanced dataset. The generated quality NAFLD dataset is subjected to training and testing of seven state of the art machine learning predictive models. The in-depth empirical analysis of testing data prediction results has shown improvement in the precision, AUC values of ROC curve indicator, which indicates significance of quality improved dataset for nafld prediction, and hence paving a way towards prediction of other diseases using our depicted step by step methodology for developing high quality prediction models in medical informatics.

Future work: the conducted experimentation needs to be extended for other diseases with real time datasets. the work can be extended with real time testing with parallel computation in real time environments. Impact of deep learning can be extended for the considered dataset.

REFERENCES

[1] M. Kivimäki, A. Bartolomucci, and I. Kawachi, “The multiple roles of life stress in metabolic disorders,” *Nature Reviews Endocrinology*, vol. 19, no. 1, pp. 10–27, 2023, doi: 10.1038/s41574-022-00746-8.

[2] N. Stefan and K. Cusi, “A global view of the interplay between non-alcoholic fatty liver disease and diabetes,” *The Lancet Diabetes and Endocrinology*, vol. 10, no. 4, pp. 284–296, Apr. 2022, doi: 10.1016/S2213-8587(22)00003-1.

[3] S. Dey *et al.*, “Human-centered explainability for life sciences, healthcare, and medical informatics,” *Patterns*, vol. 3, no. 5, p. 100493, 2022, doi: 10.1016/j.patter.2022.100493.

[4] L. Caroprese, E. Vocaturo, and E. Zumpano, “Argumentation approaches for explanaible AI in medical informatics,” *Intelligent Systems with Applications*, vol. 16, p. 200109, 2022, doi: 10.1016/j.iswa.2022.200109.

[5] R. C. Deo, “Machine learning in medicine,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[6] F. M. Delpino, K. Costa, S. R. Farias, A. D. P. Chiavegatto Filho, R. A. Arcêncio, and B. P. Nunes, “Machine learning for predicting chronic diseases: a systematic review,” *Public Health*, vol. 205, pp. 14–25, 2022, doi: 10.1016/j.puhe.2022.01.007.

[7] R. Alanazi, “Identification and prediction of chronic diseases using machine learning approach,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/2826127.

[8] C. Krittanawong *et al.*, “Machine learning prediction in cardiovascular diseases: a meta-analysis,” *Scientific Reports*, vol. 10, no. 1, p. 16057, 2020, doi: 10.1038/s41598-020-72685-1.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[10] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, “Predictive analysis of heart diseases with machine learning approaches,” *Malaysian Journal of Computer Science*, vol. 2022, no. Special Issue 1, pp. 132–148, 2022, doi: 10.22452/mjcs.sp2022no1.10.

[11] M. M. Ahsan and Z. Siddique, “Machine learning-based heart disease diagnosis: a systematic literature review,” *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 2022, doi: 10.1016/j.artmed.2022.102289.

[12] M. A. Myszczynska *et al.*, “Applications of machine learning to diagnosis and treatment of neurodegenerative diseases,” *Nature Reviews Neurology*, vol. 16, no. 8, pp. 440–456, 2020, doi: 10.1038/s41582-020-0377-8.

[13] A. M. Tăuţan, B. Ionescu, and E. Santarnecchi, “Artificial intelligence in neurodegenerative diseases: a review of available tools with a focus on machine learning techniques,” *Artificial Intelligence in Medicine*, vol. 117, p. 102081, 2021, doi: 10.1016/j.artmed.2021.102081.

[14] K. M. Kingsmore, C. E. Puglisi, A. C. Grammer, and P. E. Lipsky, “An introduction to machine learning and analysis of its use in rheumatic diseases,” *Nature Reviews Rheumatology*, vol. 17, no. 12, pp. 710–730, Dec. 2021, doi: 10.1038/s41584-021-00708-w.

[15] M. Jiang, Y. Li, C. Jiang, L. Zhao, X. Zhang, and P. E. Lipsky, “Machine learning in rheumatic diseases,” *Clinical Reviews in Allergy and Immunology*, vol. 60, no. 1, pp. 96–110, Feb. 2021, doi: 10.1007/s12016-020-08805-6.

[16] S. Nusinovici *et al.*, “Logistic regression was as good as machine learning for predicting major chronic diseases,” *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[17] G. Targher, K. E. Corey, and C. D. Byrne, “NAFLD, and cardiovascular and cardiac diseases: factors influencing risk, prediction and treatment,” *Diabetes and Metabolism*, vol. 47, no. 2, p. 101215, Mar. 2021, doi: 10.1016/j.diabet.2020.101215.

[18] G. L. H. Wong, P. C. Yuen, A. J. Ma, A. W. H. Chan, H. H. W. Leung, and V. W. S. Wong, “Artificial intelligence in prediction of non-alcoholic fatty liver disease and fibrosis,” *Journal of Gastroenterology and Hepatology (Australia)*, vol. 36, no. 3, pp. 543–550, Mar. 2021, doi: 10.1111/jgh.15385.

[19] C. C. Wu *et al.*, “Prediction of fatty liver disease using machine learning algorithms,” *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 23–29, 2019, doi: 10.1016/j.cmpb.2018.12.032.

[20] S. Qin *et al.*, “Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults,” *Scientific Reports*, vol. 13, no. 1, p. 3638, 2023, doi: 10.1038/s41598-023-30750-5.

[21] J. O’Donoghue and J. Herbert, “Data management within mHealth environments: patient sensors, mobile devices, and databases,” *Journal of Data and Information Quality*, vol. 4, no. 1, pp. 1–20, 2012, doi: 10.1145/2378016.2378021.

[22] N. Arora, S. Srivastava, R. Agarwal, V. Mehndiratta, and A. Tripathi, “Diabetes mellitus prediction using machine learning within the scope of a generic framework,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 32, no. 3, pp. 1724–1735, 2023, doi: 10.11591/IJEECS.V32.I3.PP1724-1735.

[23] A. M. Allen, T. M. Therneau, J. J. Larson, A. Coward, V. K. Somers, and P. S. Kamath, "Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: A 20 year-community study," *Hepatology*, vol. 67, no. 5, pp. 1726–1736, 2018, doi: 10.1002/hep.29546.

[24] H. Hammad Alharbi and M. Kimura, "Missing data imputation using data generated by GAN," in *ACM International Conference Proceeding Series*, 2020, pp. 73–77, doi: 10.1145/3418688.3418701.

[25] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, p. 106, 2013, doi: 10.1186/1471-2105-14-106.

BIOGRAPHIES OF AUTHORS

Nidhi Arora is currently working as associate professor Department of computer science, Kalindi College, University of Delhi. With a teaching experience of 23 years and a Ph.D. in Computer Science, University of Delhi, her primary research areas are social networks, nature inspired computing and machine learning. She has published many peers reviewed research articles in international journals of repute, book chapters, and has also presented many research papers in various international conferences of ACM, Springer and IEEE. Dr. Nidhi Arora has delivered many talks on latest research topics such as "Data Sciences", "e-content development", and "deep learning" to name a few. She has been part of many administrative assignments, committees, on the board of technical programmer committee, and acted as a reviewer to in many international journals. She can be contacted at email: nidhiarora@kalindi.du.ac.in.

Dr. Shilpa Srivastava is an associate professor and chairperson of the committee 'AI Policy recommendations' at Christ University, Bengaluru. She possesses a Ph.D. Computer Science from Uttarakhand Technical University and her areas of interest include application of soft computing in medical domain, theory of computation, algorithms, and health services. Currently she is guiding three Ph.D. Scholars and published many peers reviewed research articles in international journals of repute, one patent, book chapters, and has also presented many research papers in various international conferences of ACM, Springer and IEEE. She has delivered talks on the latest research topics such as "machine learning", "mobile learning", bring your own device", to name a few. She has also worked as CO-PI in a research project in collaboration with IIT Roorkee and Liverpool Hope University UK. She can be contacted at email: shilpa.srivastava2015@gmail.com.

Dr. Aprna Tripathi is an assistant professor in the Department of Data Science and Engineering, Manipal University Jaipur, Jaipur. She received her bachelor's degree in sciences from Kanpur University, Master's in Computer Applications from HBTI, Kanpur, M.Tech. from Banasthali University, Rajasthan and Ph.D. from NIT Allahabad, Prayagraj. With over 15 years of teaching and research experience, her scholarly contributions can be found in prestigious national and international journals and conferences, including those recognized by SCI and Scopus. Her areas of specialization include software engineering, software testing, data visualization, and data structures and algorithms. Notably, she has authored a book titled "component-based systems: estimating efforts using soft computing techniques." She can be contacted at email: aprna.tripathi@jaipur.manipal.edu.

Dr. Varuna Gupta received Ph.D. degree in Computer Science from CHRIST University, Bangalore in 2018 and a Master of Computer Applications (MCA) from AKTU, Lucknow India. She is currently working as assistant professor in the School of Sciences, CHRIST (Deemed to be University) Delhi-NCR, Ghaziabad, India. She has 17 years of teaching experience. Along with Ph.D. students Dr. Varuna has demonstrated excellent supervision abilities as a mentor in the field of machine learning and artificial intelligence for P.G. and U.G. students, applicants. Several of her works have been published in international publications and conferences. She has organized a number of university-level events, including quality development program (QIP), workshops, and faculty development programmers. She serves as a reviewer in many Scopus and WoS indexed journals. She has demonstrated remarkable efficacy in her endeavors to actively involve herself and others around him in achieving academic satisfaction. She can be contacted at email: varunagupta.cs@gmail.com.