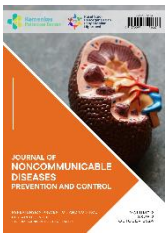
JONDPAC

JOURNAL OF NONCOMMUNICABLE DISEASES PREVENTION AND CONTROL

Volume 2, Issue 2, October 2024, pp. 43–54

ISSN 2987-1549 (Online)

DOI: https://doi.org/10.61843/jondpac.v2i2.823



Research Article

A MACHINE LEARNING BASED PREDICTION APPROACH TO NON-COMMUNICABLE DISEASES INTERVENTION

Bashir Mutebi[1], Ali Balunywa[2], Abdal Kasule[1,✉], Robert Kyeyune[2], Rogers Makubuya[1], Godfrey Mujungu[2]

[1]Makerere University Business School, Faculty of Computing and Informatics, Department of Information Systems, Kampala, Uganda, Email: bmutebi@mubs.ac.ug, akasule@mubs.ac.ug, rmakubuya@mubs.ac.ug

[2]Makerere University Business School, Faculty of Computing and Informatics, Department of Applied Computing and Information Technology, Kampala, Uganda, alibalunywa@mubs.ac.ug, rkyeyune@mubs.ac.ug, gmujungu@mubs.ac.ug

ARTICLE INFORMATION

Article history
Submitted: 10-10-2024
Revised: 11-01-2025
Accepted: 21-03-2025
Published: 25-03-2025

Keywords
Non-Communicable diseases
Machine learning algorithms
Prediction and intervention
Low and medium income countries
Uganda

ABSTRACT

This study aims to utilize machine learning techniques to predict Non-Communicable Diseases (NCDs) in Uganda, facilitating preventative actions by analyzing locally obtained data on risk factors and symptoms. A locally created dataset comprising NCDs, risk factors, and symptoms reported by medical practitioners was employed to frame NCD prediction as a classification problem. Three distinct models were developed: the first model utilized only risk factors, the second model focused solely on symptoms, and the third model integrated both risk factors and symptoms. Various machine learning classifiers, including K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes, and XGBoost, were applied to each model to assess their predictive performance. The study results indicated that KNN, was the best at predicting NCDs basing on risk factors only, while SVM was the least effective. Using symptoms to predict NCDs, ANN and Naïve Bayes emerged the best, and KNN the weakest. Using risk factors and symptoms, Random Forest was the best prediction technique while KNN was again the least effective classifier. In conclusion, this study provides valuable insights into into the comparative performance of various machine learning classifiers to model and predict NCDs using locally relevant data in Uganda. The findings underscore the importance of accurately predicting NCDs at early stages, enabling medical personnel to intervene and offer preventive treatments to high-risk individuals. The identification of the most effective classifiers paves the way for future research and implementation initiatives in low- and middle-income countries.

Kata Kunci
Penyakit tidak menular
Algoritma pembelajaran mesin
Prediksi dan intervensi
Negara berpenghasilan rendah dan menengah
Uganda

ABSTRAK

Penelitian ini bertujuan memanfaatkan teknik pembelajaran mesin untuk memprediksi penyakit tidak menular (PTM) di Uganda, memfasilitasi tindakan pencegahan dengan menganalisis data yang diperoleh secara lokal mengenai faktor risiko dan gejalanya. Kumpulan data lokal yang terdiri atas PTM, faktor risiko, dan gejala yang dilaporkan oleh praktisi medis digunakan untuk menyusun prediksi PTM sebagai klasifikasi masalah. Tiga model berbeda telah dikembangkan: model pertama hanya memanfaatkan faktor risiko, model kedua hanya berfokus pada gejala, dan model ketiga mengintegrasikan faktor risiko dan gejala. Berbagai pengklasifikasi pembelajaran mesin, termasuk KNN, Random Forest, SVM, ANN, Naïve Bayes, dan XGBoost, diterapkan pada setiap model untuk menilai performa prediktifnya. Hasilnya menunjukkan bahwa KNN, Random Forest, dan ANN adalah pengklasifikasi dengan kinerja terbaik untuk Model 1 (hanya faktor risiko), sedangkan SVM adalah yang paling tidak efektif. Pada Model 2 (hanya gejala), ANN dan Naïve Bayes muncul sebagai pengklasifikasi terbaik, diikuti oleh SVM dan XGBoost, dan KNN menjadi yang terlemah. Untuk Model 3 (input gabungan), Random Forest berkinerja terbaik, digantikan oleh XGBoost, sementara KNN sekali lagi merupakan pengklasifikasi yang paling tidak efektif. Studi ini memberikan wawasan berharga mengenai pemodelan prediktif PTM dengan menggunakan data lokal yang relevan di Uganda. Temuan ini menggarisbawahi pentingnya memprediksi PTM secara akurat pada tahap awal, sehingga memungkinkan tenaga medis untuk melakukan intervensi dan menawarkan perawatan pencegahan kepada individu yang berisiko tinggi. Identifikasi pengklasifikasi yang paling efektif membuka jalan bagi penelitian di masa depan dan inisiatif implementasi di negara-negara berpenghasilan rendah dan menengah.

✉ ***Corresponding Author:***

Abdal Kasule
Makerere University Business School, Faculty of Computing and Informatics, Department of Information Systems, Kampala, Uganda
Email: akasule@mubs.ac.ug

Citation:

Mutebi, B., Balunywa, A., Kasule, A., Kyeyune, R., Makubuya, R., & Mujungu, G. (2025). A Machine Learning Based Prediction Approach To Non-Communicable Diseases Intervention. *Journal of Noncommunicable Diseases Prevention and Control*. *2*(2): 43-54.

INTRODUCTION

The global incidence of non-communicable diseases (NCDs) is on the rise, a trend that is particularly evident in developing nations. NCDs, which include cardiovascular diseases, diabetes, cancer, and chronic respiratory diseases, account for a higher rate of morbidity and mortality than all other causes of illness combined. Kusolo *et al*. (2024), Natukwatsa *et al*. (2021). As such, they represent a significant health and developmental challenge for humanity in the twenty-first century. According to estimates provided by the World Health Organization (WHO), non-communicable diseases (NCDs) account for approximately 41 million fatalities annually, representing 74% of all global deaths. Furthermore, it is reported that 17 million individuals succumb to NCDs before reaching the age of 70 each year. Notably, a significant proportion of these premature deaths, amounting to 86%, occur in low- and middle-income countries. Additionally, it is important to highlight that 77% of all deaths attributed to NCDs take place within these same socioeconomic contexts (WHO, Fact Sheet, September 2023).

Non-communicable diseases (NCDs) account for approximately 70% of all deaths in developing countries, with nearly half of these fatalities occurring in individuals under the age of 70. Projections indicate that the global burden of NCDs will rise by 17% over the next decade, with an even more pronounced increase of 27% anticipated in the African region. It is important to note that developing countries and regions continue to grapple with the repercussions of infectious diseases, while simultaneously experiencing a rapid escalation in the burden of NCDs. This dual challenge is particularly pronounced in low- and middle-income countries (LMICs), where NCDs have emerged as significant impediments to sustained economic development and progress (Wang & Wang, 2020). Numerous patients diagnosed with non-communicable diseases (NCDs) often present with pronounced symptoms only upon reaching the intermediate or terminal stages of their conditions. At this juncture, the diseases may become nearly irreversible, particularly following the development of substantial lesions. These attributes underscore the critical importance of early detection in the prevention and management of NCDs (Wu *et al*., 2022).

Similar to other Low and Middle-Income Countries (LMICs), Uganda is experiencing a significant increase in the burden of major non-communicable diseases (NCDs), which can be attributed to several interrelated factors. These include rapid population growth, urbanization, nutritional transitions, and both indoor and outdoor air pollution (Rogers *et al*., 2018).

Uganda faces a growing burden of non-communicable diseases (NCDs). While a specific national figure on NCD-attributable deaths since 2014 is unavailable, The Ugandan Ministry of Health (2021) acknowledges this burden. Earlier data highlights the number of high blood pressure cases in outpatient departments increased from 60,000 in 2012-2013 to 85,000 in 2015-2016, and diabetes cases rose by 7% during the same period (Uganda Ministry of Health, 2021). These statistics underscore the urgency for comprehensive public health interventions and improved monitoring to address this growing challenge. The World Health Organization (WHO) also reports a high age-standardized mortality rate in 2021. This rate reached 709 per 100,000 for males and 506 per 100,000 for females for four major NCDs: cardiovascular diseases, chronic respiratory diseases, cancers, and diabetes (World Health Organization, African Region, 2023). As of 2022, non-communicable diseases (NCDs) in Uganda accounted for approximately 40% of total deaths, up from 33% in 2016. Among these NCDs, cardiovascular diseases contributed to 12%, cancers to 10%, chronic respiratory diseases to 3%, diabetes remained at 2%, and other NCDs made up the remaining 13%. These changes, highlighted by the World Health Organization (2023), Ministry of Health Uganda (2021) and WHO Regional Office for Africa (2023), point to a shifting health landscape in Uganda. The increasing burden from NCDs necessitates targeted health interventions alongside continued efforts to manage communicable diseases.

At present, global awareness among patients regarding early-stage non-communicable diseases (NCDs) is notably limited and exhibits significant variability across different diseases and geographical regions. Consequently, in order to enhance the management of NCDs, it is imperative to establish effective early prediction methodologies that can elevate patient awareness concerning undiagnosed conditions. Recent advancements in machine learning methodologies, coupled with the extensive accumulation of Electronic Health Records (EHR) data, have rendered the early prediction of non-communicable diseases (NCDs) based on real-world data a feasible endeavor, thereby attracting significant scholarly interest. Numerous studies have concentrated on the early identification of conditions such as diabetes mellitus (DM), hypertension (HTN), cardiovascular diseases, and other NCDs, as well as on the simultaneous prediction of various chronic diseases. Research in this domain can be categorized into two primary types: static (one-time) prediction and temporal

prediction, the latter of which employs sequential data for analysis (Wu *et al*., 2022). However, optimal and sustainable implementation strategies for such interventions within the LMIC context require locally led and conducted research – a capacity that is currently lacking (Engelgau *et al*., 2018). Therefore, there is a critical need for robust predictive and preventive intervention mechanisms based on locally available data. This study aims to address this gap by using machine learning approaches on locally collected Ugandan data to build a prediction and preventive model for NCD intervention.

RELATED WORK

Non-communicable diseases (NCDs) represent a significant public health challenge globally, with an acute impact on developing countries like Uganda. These diseases, including cardiovascular diseases, diabetes, cancers, and chronic respiratory diseases, are major contributors to morbidity and mortality globally (World Health Organization, 2021). NCDs are primarily driven by; modifiable behavioral risk factors such as poor diet, lack of physical activity, tobacco use, and excessive alcohol consumption (Wensonga *et al*., 2016); metabolic risk factors such as raised blood pressure, overweight/obesity, hyperglycemia (high blood glucose levels) and hyperlipidemia (high levels of fat in the blood) and several environmental risk factors such as air pollution (Davagdorj *et al*., 2021).

In Uganda, NCDs accounted for approximately 40% of total deaths as of 2022, up from 33% in 2016 (World Health Organization, 2023; Ministry of Health Uganda, 2021). Cardiovascular diseases contributed 12%, cancers 10%, chronic respiratory diseases 3%, diabetes 2%, and other NCDs made up the remaining 13%. This shift highlights a growing health burden that requires targeted interventions alongside efforts to manage communicable diseases (World Health Organization, 2023; Ministry of Health Uganda, 2021). This rapid increase in NCD necessitates effective intervention strategies to manage and mitigate their impact on the population's health.

Machine learning (ML) has emerged as a powerful tool for NCD prediction and management in healthcare. ML algorithms can analyze complex datasets to identify patterns and predict disease onset with high accuracy. Various studies have demonstrated the efficacy of ML in predicting conditions such as diabetes, cardiovascular diseases, and other NCDs using diverse data sources. For instance, Pranto *et al*. (2020) focused on predicting diabetes among female patients using a range of machine learning techniques. Tasin *et al*. (2022) developed a system for predicting diabetes using machine learning and explainable AI techniques. Their model employed the Pima Indian diabetes dataset and a private dataset from Bangladesh, achieving high accuracy with ensemble methods like XGBoost and ADASYN. The use of explainable AI techniques, such as SHAP (Shapley Additive Explanations), provided insights into the model's decision-making process, enhancing its clinical applicability. Ferdousi *et al*. (2021) introduced an innovative framework for health cyber-physical systems (CPS) that employs machine learning techniques to tackle the challenges associated with the efficient processing of wearable Internet of Things (IoT) sensor data. This framework aims to facilitate early risk prediction of diabetes. Experiments using several machine learning algorithms showed that the approach was effcetcive and achieved an accuracy of 94% from the Random Tree algorithm in the prediction of diabetes at an early stage. Focusing on cardiovascular disease (CVD) prediction, Subramani *et al*. (2023) employed Random Forest and Support Vector Machines on extensive patient datasets, demonstrating significant predictive performance. This study aligns with the potential of ML for identifying individuals at high risk of CVD, enabling earlier interventions. This study underscores the potential of ML in identifying high-risk individuals and enabling timely interventions to mitigate the impact of CVD. Marbaniang *et al*., (2020) investiagetd several supervised machine learning classifiers to predict CVD. Among the six algorithms examined—namely K-Nearest Neighbors, Naïve Bayes, Decision Trees, Random Forest, Support Vector Machine, and Linear Discriminant Analysis—the K-Nearest Neighbor Classifier demonstrates the highest predictive accuracy. Mukherjee *et al*. (2021) implemented ML algorithms, including logistic regression and decision trees, to predict hypertension risk based on routine health data. This emphasizes the potential of ML for leveraging existing data sources in resource-limited settings like Uganda for early NCD detection. Islam *et al*. (2023) used Decision trees and Random Forests to predict hypertension, there was an improved accuracy and robustness in prediction tasks. Davagdorj *et al*. (2021) developed an Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction using Support Vector Machines (SVM). Jackins *et al*. (2021) demonstrated the effectiveness of Neural Networks and Deep Learning techniques in handling large datasets and capturing non-linear relationships. Their study highlighted the utility of models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for predicting NCDs,

emphasizing their superior ability to extract intricate features from data. Ensemble methods like XGBoost and Gradient Boosting Machines (GBM), which combine multiple models to enhance prediction accuracy, were utilized by Tasin *et al*. (2022) to predict diabetes on imbalanced medical datasets.

In the context of Uganda, Explainable AI (XAI) that provides insights into how models make decisions, enhancing trust and transparency in clinical settings (Tasin *et al*., 2022) could facilitate the adoption of ML-based decision support systems by providing transparent and understandable predictions to healthcare providers. This is particularly important in environments where there may be skepticism or limited familiarity with advanced technological tools.

MATERIALS AND METHODS

Problem Formulation

The NCD prediction problem was formulated as a classification problem based on inputs and outputs. The inputs were risk factors and symptoms and output was presence of a given NCD that corresponds to the given risk factors, symptoms or both. Let $X = [X_1, X_2, \ldots\ldots, X_n]$ be the feature matrix containing risk factors and symptoms, where n is the number of samples, with each sample X_i containing d features, $X_i = [x_{i,1}, x_{i,2}, \ldots\ldots, x_{i,d}]$. Let $y = [y_i, y_2, \ldots\ldots, y_n]$ be the output vector indicating the presence or absence of a particular NCD in a target sample i. The objective is to find a function $f: R^d \rightarrow \{0,1\}$ that maps the feature vector x to the output y. That is

$$y = f(x) \qquad \text{eqn (1)}$$

Data Collection

Enough preprocessing and reliable datasets are needed for NCD prediction to be effective. Electronic health records (EHRs), public health surveys, and particular illness registries are examples of common data sources. Online questionnaires were used in this study to gather data from medical practitioners who serve patients with noncommunicable diseases. Healthcare experts' reports of NCDs, risk factors, and symptoms made up the locally generated dataset. **Figures 1, 2, and 3** show the distribution of NCDs, risk factors, and symptoms, respectively, based on the locally generated dataset. The NCDs, risk factors, and symptoms for each NCD as determined by the healthcare professionals were contained in the dataset for this study, in contrast to traditional datasets that includes values for parameters like age, weight, body mass index, etc.

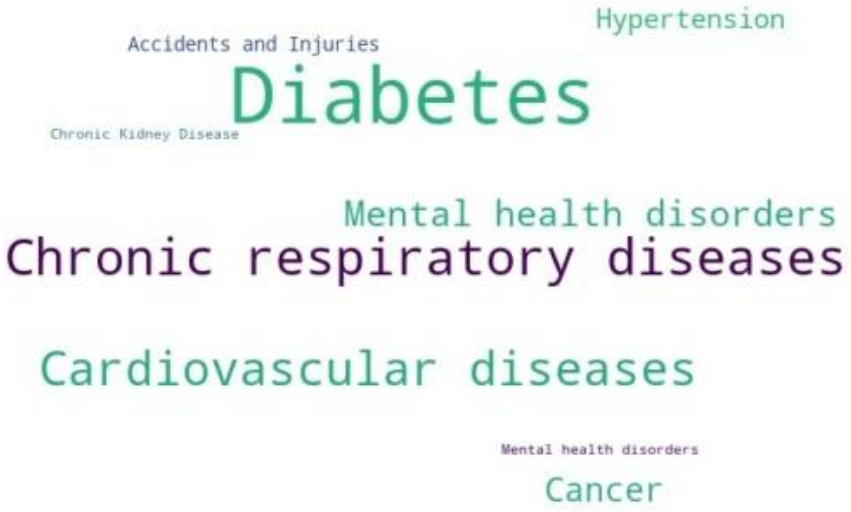


Figure 1 NCDs in the dataset

Figure 1 illustrates how healthcare providers mostly deal with patients who have diabetes, followed by those who have chronic respiratory conditions, cardiovascular disorders, and cancer, in that order. **Figure 2** shows that among other risk factors, alcohol drinking, poor diet, family history, inactivity, and tobacco use were the most prevalent ones.

Figure 2 Risk factors in the dataset

Figure 3 illustrates the main symptoms in the dataset, which included exhaustion or weakness, irregular heart rate, high blood pressure, and raised blood sugar levels. This agrees with Wesonga *et al.* (2016)'s study in which high frequency of tobacco smoking, less than five servings of fruit and vegetables per day, low physical activity levels and raised blood pressure.

Figure 3 Symptoms in the dataset

The Procedure

The process is depicted in **Figure 4**. This process is based on the standard machine learning methodology. Preprocessing the data, choosing features, splitting the dataset into training and testing datasets, classifying the data, and finally generating the predictions are the first phases in the process. The data pre-processing include eliminating null values and incomplete data. The data in the dataset was textual and was converted into numerical data in order to be used for the classification tasks. NCDs were considered a categorical target variable. The label encoder function was used to convert this into numerical numbers. Using the get_dummies pandas function, the risk factors and symptoms were converted into numerical variables (1s and 0s). 179 characteristics were obtained from the dataset following this modification. We utilized a ranking algorithm to choose the best attributes from this feature set to include in the model. Removing redundant and unnecessary features and keeping just those that are pertinent to the learning and performance of the algorithms is the aim of feature selection (Mladenić, 2011).

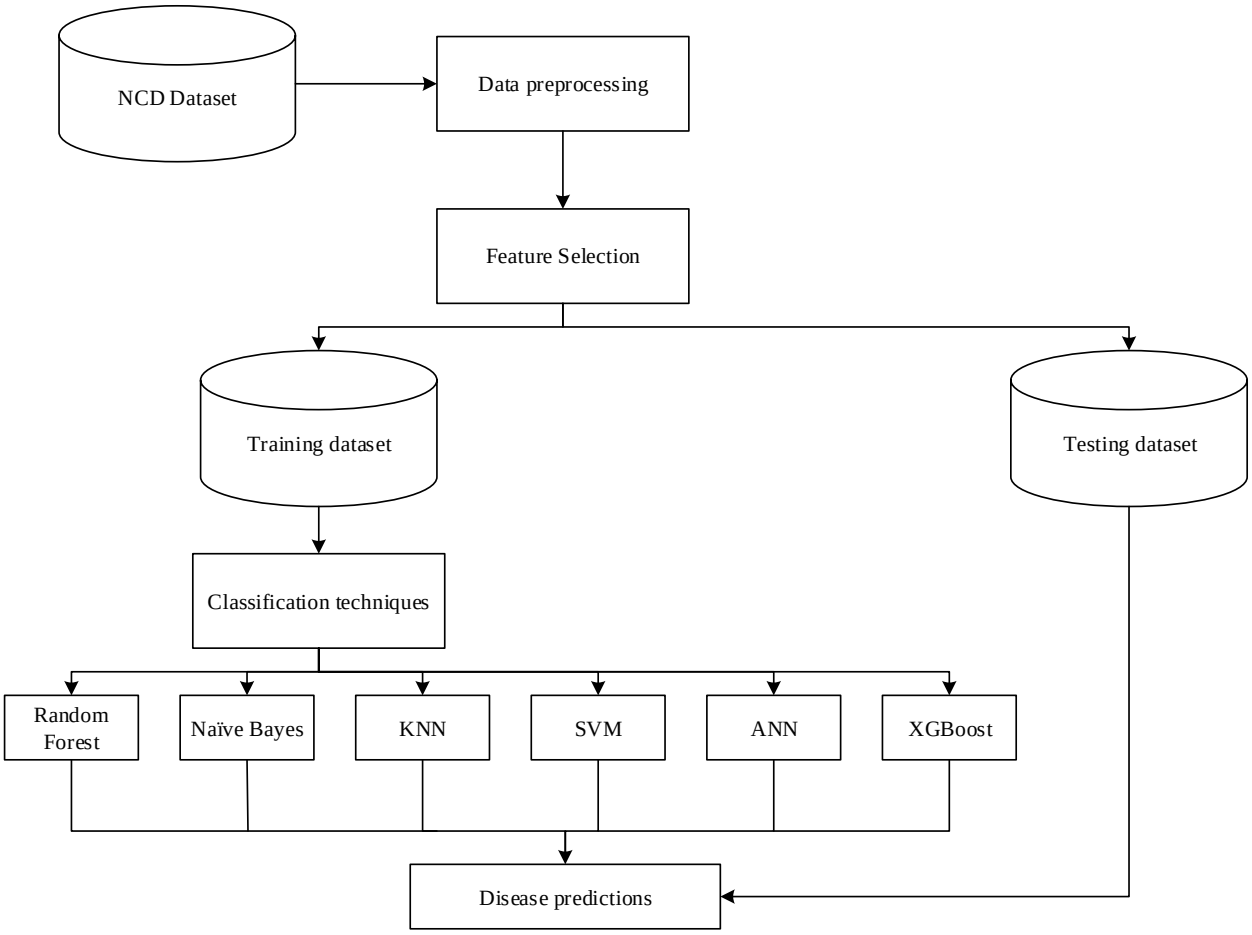


Figure 4 Procedure for the NCD predictive models

Classification is a supervised machine learning technique that uses predictor and target variables. NCDs were the focus of this study, and the predictor variables were risk factors and symptoms. A total of three distinct predictive models were created: the first one used risk factors as the only inputs (predictors), the second one used symptoms alone, and the third one combined risk factors and symptoms as predictors. Entries 2, 3, and 4 display the three models.

$$Y = f(X_1), X_1 = [x_{1,1}, x_{1,2}, \ldots., x_{1,d}] \text{ eqn (2).}$$
$$Y = f(X_2), X_2 = [x_{2,1}, x_{2,2}, \ldots., x_{2,d}] \text{ eqn (3).}$$
$$Y = f(X_3), X_3 = [x_{3,1}, x_{3,2}, \ldots., x_{3,d}] \text{ eqn (4).}$$

Where X_1, X_2, and X_3 are input vectors containing risk factors only, symptoms only and both risk factors and symptoms. Y is a vector containing the NCDs. All the models were trained and evaluated using six supervised machine learning algorithms. These algorithms included Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost, and Artificial Neural Network (ANN).

Evaluation Metrics

Typically, classification problems involve classifying a target correctly or incorrectly. The classification outputs can be categorized into four distinct groups: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). True Positives refer to the instances in which the model accurately identifies samples belonging to the positive class. Conversely, True Negatives denote the instances where the model correctly identifies samples belonging to the negative class. False Positives represent the instances in which the model incorrectly classifies negative class samples as positive. Finally, False Negatives indicate the instances where the model erroneously predicts positive class samples as negative. The study thus used common classification metrics such as Accuracy, Precision, Recall and F1-Score, to determine the predictive effectiveness of the models. These metrics are always represented as percentages or values between 0 and 1. The higher the percentage or closer to 1 the value of the metric is, the better the classifier.

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. In other words, Accuracy is the number of correct predictions divided by the total number of predictions (Gang *et al*., 2020).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{eqn (5)}$$

The precision metric concentrates on FP and is defined as the percentage of accurately identified positive cases. When your precision score is close to 1, it means that your model accurately distinguishes between the proper and wrong labeling of targets in the dataset and that it didn't miss any true positives. A low precision score (<0.5) indicates a high rate of false positives in your classifier, which may be the result of misaligned class or untuned hyper-parameters in the model.

$$Precision = \frac{TP}{TP+FP} \qquad \text{eqn (6)}$$

Recall is the actual positive cases that are correctly identified. A Recall is essentially the ratio of true positives to all the positives in ground truth.

$$Recall = \frac{TP}{TP+FN} \qquad \text{eqn (7)}$$

F1-score is a harmonic mean of precision and recall, that provides a single metric that balances both the precision and the recall.

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad \text{eqn (8)}$$

A low F1 score indicates that the model is either failing to predict the target variable when it is present or predicting it as present when it is not. A higher F1 score, on the other hand, indicates a better balance between precision and recall, making it a reliable measure of a model's effectiveness.

RESULTS AND DISCUSSION

The **Table 1, 2, and 3** show the metrics for each of the classifiers for the three models.

Table 1 Performance metrics for model 1 using only risk factors as inputs

Classifier	Accuracy(%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	54.55	42.86	44.29	39.05
Naïve Bayes	36.36	25.00	15.71	19.05
SVM	27.27	7.14	10.00	8.33
XGBoost	27.27	22.22	15.00	17.17
ANN	54.55	42.86	44.29	39.05
KNN	63.64	42.86	40.00	41.27

From **Table 1**, the best classifier is KNN with an accuracy of 63.64%, precision of 42.86%, Recall of 40.00%, and an F1-Score of 41.27%. This is followed by Random Forest and ANN with similar values for all the metrics. The worst classifier is SVM with an accuracy of 27.27%, Precision of 7.14%, Recall of 10.00% and F1-Score of 8.33%.

In **Table 2**, ANN and Naïve Bayes are the best classifiers with an accuracy of 90.91%, and a Recall of 83.33%, however, ANN has a precision of 75.00%, and F1-Score of 77.78%. Naïve Bayes on the other hand has a precision of 80.56%, and an F1-Score of 81.82. This is followed by Random Forest, SVM and XGBoost. The worst classifier is KNN with an accuracy of 27.27%, precision of 27.08%, Recall of 36.67% and F1-Score of 20.37%. This is contrary to the results of model 1 where KNN is the best classifier.

Table 2 Performance metrics for model 2 using only symptoms as inputs

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	81.82	72.22	75.00	70.71
Naïve Bayes	90.91	80.56	83.33	81.82
SVM	63.64	28.57	50.00	36.11
XGBoost	54.55	42.86	44.29	39.29
ANN	90.91	75.00	83.33	77.78
KNN	27.27	27.08	36.67	20.37

In **Table 3**, Random Forest, SVM and ANN have an accuracy of 81.82%. However, Random Forest has better precision (88.89%), Recall (88.33%) and F1-Score (84.26%). This makes Random Forest the best classifier when both risk factors and symptoms are used as predictors. The superior performance of Random Forest can be attributed to its ensemble learning approach, which combines the predictions of multiple decision trees to enhance overall accuracy and reduce the risk of overfitting. The model's ability to handle a mix of categorical and continuous variables, along with its robustness to noise in the data, further contributes to its effectiveness in this context. Following Random Forest, XGBoost demonstrated commendable performance, although it did not match the precision and recall metrics of Random Forest. XGBoost is recognized for its gradient boosting framework that optimizes the classification process through an iterative approach. XGBoost typically excels in scenarios with complex relationships in the data, and its ability to manage missing values and prevent overfitting makes it a strong contender in various classification tasks. The worst classifier is KNN with an accuracy of 45.45%, precision of 41.67%, Recall of 48.33%, and F1-Score of 33.97%. The poor performance of KNN can be attributed to its reliance on distance metrics to classify instances, which can be problematic in high-dimensional spaces or when the data contains noise.

Table 3 Performance metrics for model 3 using both risk factors and symptoms as inputs

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	81.82	88.89	88.33	84.26
Naïve Bayes	63.64	38.89	41.67	37.37
SVM	81.82	50.00	66.67	55.56
XGBoost	72.73	40.48	44.29	36.90
ANN	81.82	75.00	83.33	77.78
KNN	45.45	41.67	48.33	33.97

The graph in **Figure 5**, is a comparative visualization of the aggregated metrics of the different classifiers. From this visualization, it can be noted that overall, ANN is the best classifier, followed by Random Forest, Naïve Bayes, SVM, XGBoost and lastly KNN.

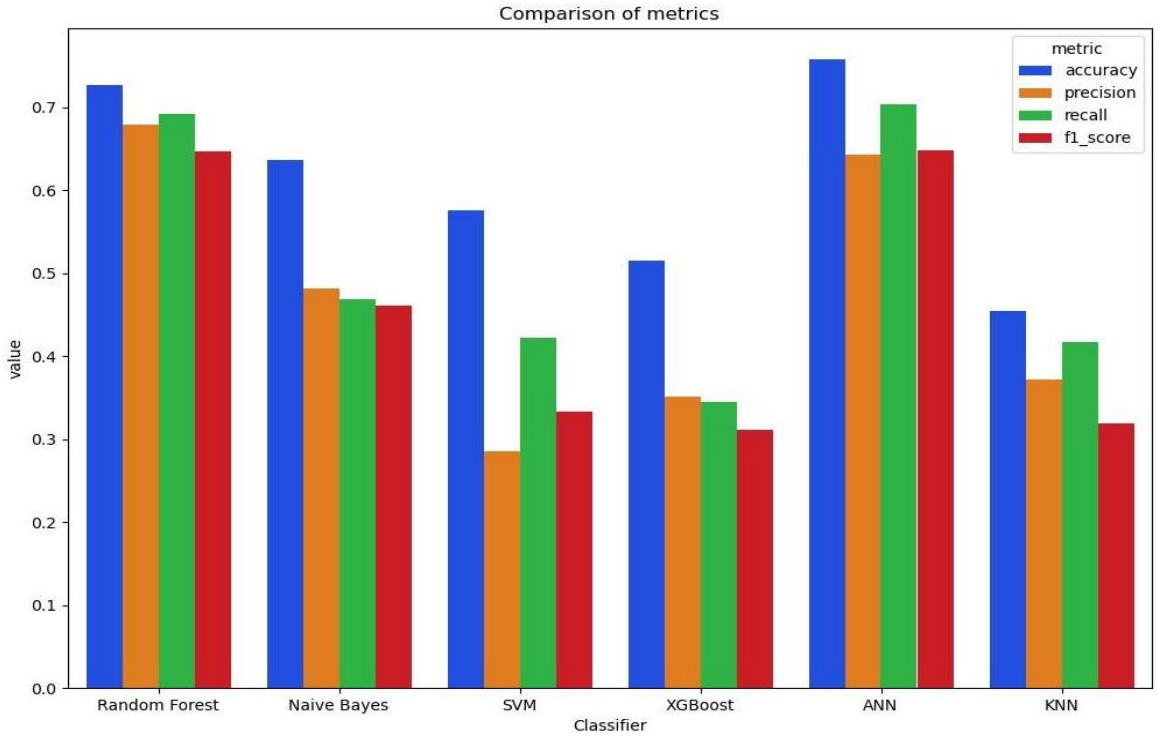


Figure 5 Comparison of the classifiers by aggregated metrics

The goal of this study was to create a machine learning-based prediction method for the intervention of non-communicable diseases (NCDs). We assessed how well different classifiers performed in forecasting non-communicable diseases (NCDs) using distinct sets of variables, such as risk factors and symptoms. Our study's findings show that the classifiers' performance differed according on the predictor set that was employed. K-nearest neighbors (KNN) was the most accurate classifier in model 1, which only took risk variables into account as predictors. However, due to its poorer precision and recall, KNN, like all other classifiers, had trouble correctly identifying positive cases. On the other hand, Random Forest and Artificial Neural Network (ANN) performed relatively well in model 1. These results indicate that Random Forest and ANN were able to achieve a good balance between accuracy and other performance metrics.

In model 2, the classifiers' performance shifted when symptoms were taken into account as predictors. According to these findings, both ANN and Naïve Bayes were able to obtain high recall and accuracy; however, Naïve Bayes outperformed ANN in terms of F1-Score and precision. In model 3, Random Forest

outperformed the other classifiers as well. This shows that Random Forest performed better as a result of its ability to use risk variables and symptoms as predictors. In every model, SVM and XGBoost performed comparatively worse. These findings suggest that SVM had difficulty correctly predicting NCDs using the provided variables. Finally, KNN's performance varied throughout the models. In model 1, it was the top classifier; however, in models 2 and 3, it did not perform well. This implies that when symptoms are included as predictors, KNN might not be appropriate for NCD prediction.

Overall, the findings indicate that ANN, Random Forest, and Naïve Bayes were the top-performing classifiers in predicting NCDs. Among these, Random Forest exhibited the best overall performance when both risk factors and symptoms were considered as predictors. These results highlight the importance of selecting appropriate classifiers and predictor sets when developing machine learning-based approaches for NCDs intervention.

CONCLUSION

The results of this study have important implications for NCD intervention strategies. If NCDs are predicted NCDs at an early stage, healthcare professionals can intervene and provide preventive measures to individuals at high risk. This can lead to better health outcomes and potentially reduce the burden of NCDs on healthcare systems. The identification of top-performing classifiers, such as ANN, Random Forest, and Naïve Bayes, can guide future research and implementation efforts in LMICs.

It is crucial to remember that the predictor set that was employed affected how well the classifiers performed. The study concludes that future prediction models should take symptoms into account as they are critical in the prediction of NCDs. Although the focus of this study was on NCD prediction in Uganda, other LMICs with comparable healthcare systems and data availability may find value in the findings. In areas where resources are limited, the application of machine learning techniques offers considerable promise for enhancing NCD intervention tactics. It is crucial to keep in mind, nevertheless, that the classifiers' performance could differ based on the particular dataset and demographics. Therefore, before using the prediction models in real-world situations, more validation and improvement are required.

The study found out that using a combination of risk factors and symptoms yielded better results as compared to using only risk factors and symptoms only. This study also demonstrates the feasibility and effectiveness of machine learning-based approaches for NCD prediction and intervention. The findings highlight the importance of selecting appropriate classifiers and predictor sets. These results contribute to the growing body of literature on NCD prediction and provide valuable insights for healthcare professionals and policymakers in LMICs. Future research should focus on validating and refining the prediction models in diverse populations and healthcare settings, ultimately leading to improved NCD intervention strategies and better health outcomes for individuals at risk.

CONFLICT OF INTEREST

No conflict of interest or common interest has been declared by the authors.

ACKNOWLEDGEMENTS

The researchers acknowledge the financial support for the research, from the Faculty of Computing and Informatics, Makerere University Business School.

REFERENCES

Davagdorj, K., Theera-Umpon, N., Bae, J.-W., Pham, V.-H., & Ryu, K. H. (2021). Explainable artificial intelligence based framework for non-communicable diseases prediction. IEEE Access, 9, 123672-123679. https://doi.org/10.1109/ACCESS.2021.3110336

Engelgau, M. M., Rosenthal, J. P., Newsome, B. J., Price, L., Belis, D., & Mensah, G. A. (2018). Noncommunicable diseases in low-and middle-income countries: a strategic approach to develop a global implementation research workforce. Global heart, 13(2), 131-137.

Ferdousi, R., Hossain, M. A., & El Saddik, A. (2021). Early-stage risk prediction of non-communicable disease using machine learning in health CPS. IEEE Access, 9, 96823-96837.

Islam, M. M., Alam, M. J., Maniruzzaman, M., Ahmed, N. A. M. F., Ali, M. S., Rahman, M. J., & Roy, D. C. (2023). Predicting the risk of hypertension using machine learning algorithms: A cross-sectional study in Ethiopia. PLoS One, 18(8), e0289613. https://doi.org/10.1371/journal.pone.0289613

Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y.: AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J. Supercomput. 77, 5198–5219 (2021)

Kusolo, R., Mutungi, G., Mbuliro, M., Kajjura, R., Wesonga, R., Bahendeka, S. K., & Guwatudde, D. (2024). Changes in the prevalence of the common risk factors for non-communicable diseases in Uganda between 2014 and 2023: Informed by nationally representative cross-sectional surveys. medRxiv. https://doi.org/10.1101/2024.09.04.24313080

Marbaniang, I. A., Choudhury, N. A., & Moulik, S. (2020). Cardiovascular disease (CVD) prediction using machine learning algorithms. In 2020 IEEE 17th India council international conference (INDICON) (pp. 1-6). IEEE.

Mladenić, D. (2011). Feature Selection in Text Mining. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_307

Ministry of Health Uganda. (2021). The status of non-communicable diseases (NCDs) health care in Uganda. Retrieved from https://www.health.go.ug

Ministry of Health Uganda. (2014). Non-Communicable Disease Risk Factor Baseline Survey. Retrieved from Ministry of Health Uganda (https://www.health.go.ug/cause/non-communicable-disease-risk-factor-baseline-survey/)

Natukwatsa, D., Wosu, A. C., Ndyomugyenyi, D. B., Waibi, M., & Kajungu, D. (2021). An assessment of non-communicable disease mortality among adults in Eastern Uganda, 2010–2016. PLOS ONE, 16(3), e0248966. https://doi.org/10.1371/journal.pone.0248966

Pranto, B., Paul, S., Rifat, M. R., & Barua, S. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information, 11(1), 1-20. https://doi.org/10.3390/info11010020

Rogers, H. E., Akiteng, A. R., Mutungi, G., Ettinger, A. S., & Schwartz, J. I. (2018). Capacity of Ugandan public sector health facilities to prevent and control non-communicable diseases: an assessment based upon WHO-PEN standards. BMC health services research, 18, 1-13.

Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine, 10*, 1150933. https://doi.org/10.3389/fmed.2023.1150933

Tasin, I., Ullah, N., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1-2), 1-10.

Wang, Y., & Wang, J. (2020). Modelling and prediction of global non-communicable diseases. BMC public health, 20, 1-13.

Wesonga, R., Guwatudde, D., Bahendeka, S. K., Mutungi, G., Nabugoomu, F., & Muwonge, J. (2016). Burden of cumulative risk factors associated with non-communicable diseases among adults in Uganda: Evidence from a national baseline survey. International Journal for Equity in Health, 15(1), 195. https://doi.org/10.1186/s12939-016-0486-6

WHO Regional Office for Africa. (2023). Uganda. Retrieved from https://www.afro.who.int/sites/default/files/2023-08/Uganda.pdf.

WHO. (2024). Burden of non-communicable diseases on the rise. Retrieved from [WHO Regional Office for Africa] https://www.afro.who.int.

WHO, FACT SHEET, SEPTEMBER 2023. Retrieved from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-

diseases#:~:text=Noncommunicable%20diseases%20(NCDs)%20kill%2041,%2D%20and%20middle%2Dincome%20countries, 12th February 2024.

World Health Organization. (2021). Non-Communicable Diseases. Retrieved from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

World Health Organization. (2023). Surveillance of non-communicable diseases. Retrieved from https://www.who.int

Wu, C., Zhou, T., Tian, Y., Wu, J., Li, J., & Liu, Z. (2022). A method for the early prediction of chronic diseases based on short sequential medical data. Artificial Intelligence in Medicine, 127, 102262.