International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Machine Learning Algorithms for Early Detection of Chronic Diseases in the Medicare Population

Ginoop Chennekkattu Markose

Engineer Lead Sr, EDA- Risk and Quality Digital Solutions
Elevance Health Inc, Richmond, Virginia, United States

Abstract: ***Diabetes, cardiovascular disease, and COPD are characteristic examples of chronic diseases that contribute greatly to the general healthcare cost, especially for the Medicare population. The best management of these conditions and the resulting morbidity and mortality is dependent on early screening for these conditions. Artificial Intelligence (AI) has become significant in the healthcare sector owing to the possibility of processing a large amount of data so as to compare with earlier data and infer the early stages of chronic diseases. This paper identifies and discusses several ML algorithms that can be used for the early detection of conditions in the Medicare population. As for the algorithms, the work is devoted to the aspects of data preprocessing, model selection, and evaluation measurements used in clinical practice. In this paper, we compare algorithms such as the SVM, RF, and NN using a Medicare dataset. Hence, the study reveals that embracing the use of modern technologies such as ML could help Medicare easily detect chronic diseases, leading to more effective treatment and or management of such patients, as well as rational use of available resources. Lastly, the paper includes points of concern, possibilities of having bias, and directions for future work when using ML in this area.***

Keywords: Chronic diseases, Machine Learning, Medicare, Early detection, Support Vector Machines, Random Forest, Neural Networks, Healthcare.

1. Introduction

Diseases such as diabetes, heart disease, and Chronic Obstructive Pulmonary Disease affect a large population in the world, have high mortality, long-time disability rates, and bear a high health cost to society. The USA observes that over 60 percent of the population is affected by chronic illnesses, a fact that is reflected in the fact that chronic diseases account for more than 90% of all health costs in the country. This is evident in the Medicare program through which over 64 million older Americans are enrolled, and the program is constantly under pressure thanks to the current high incidence of chronic diseases [1].

1.1 Early Detection Importance

Screening is instrumental in the early identification of diseases, the creation of a healthier society and timely intervention and treatment. Present techniques are frequently inadequate, as early diagnostics are not achieved by the common manifestations or through ordinary tests and examinations. Artificial intelligence, especially the subset called Machine learning (ML), is showing much promise as an enabler tool in disease diagnosis. By using training data such as patients' electronic health records, vital sign data from wearable devices and lab results, the ML algorithms can look for early markers that the disease is on set.

1.2 Machine learning in health care

ML has the characteristics of working with multiple dimensions, and its performance can only get better with time as it receives more data on the particular problem. They used this adaptive learning process when new research and patient data were released constantly in the field of healthcare. Three, it emphasizes that besides accurately identifying the earliest signs of the disease, ML can help in risk assessment and individualizing the approach, thanks to which medical practitioners can focus on patients who are at high risk of mortality and adjust the treatment method depending on the patient's traits.

Despite the limitations that come with the use of ML in healthcare, including issues to do with data quality, as well as the bias that could be inherent in the algorithms, the benefits of using the technology in an attempt to diagnose and manage chronic diseases cannot be overstated. The implementation of an ML-based clinical decision support system is the key to changing the paradigms of chronic disease detection and management for the sake of obtaining improved patient and system outcomes.

Generally, the integration of the ML into clinic practice care could be a major paradigm shift in how chronic illness is managed thereby leading to better outcomes for patients and the entire healthcare system.

2. Literature Survey

The use of Machine Learning (ML) algorithms in chronic disease diagnosis has been increasingly popularized in the recent past especially for the Medicare population group. This paper aims to review the literature on the application of ML algorithms in the early diagnosis of chronic diseases, with particular emphasis on the elderly since they are the most affected. The review includes information on the rate of chronic diseases in Medicare beneficiaries, conventional diagnostic techniques, the introduction of ML in healthcare, concrete types of ML used in the identification of chronic diseases, and the difficulties in implementing such technologies in practice.

2.1 Chronic Diseases in the Medicare Population

The most common illnesses are chronic diseases, including diabetes, cardiovascular diseases, as well as COPD, and most of the patients are from the elderly group under

Medicare. These conditions generate a large portion of health care use and expenditure in Medicare.

For instance, cardiovascular diseases are the leading cause of hospitalization among Medicare beneficiaries, accounting for millions of hospital stays each year. Diabetes, another common condition, affects nearly one-third of Medicare beneficiaries, leading to significant complications such as kidney disease, neuropathy, and cardiovascular events if left unmanaged. COPD is also prevalent, particularly among elderly smokers, and is a leading cause of morbidity and mortality in this population.

The late detection of these diseases often results in advanced disease stages at diagnosis, leading to more severe health outcomes and higher treatment costs. This underscores the urgent need for early detection mechanisms that can identify these conditions before they become symptomatic.

2.2. Traditional Diagnostic Methods

Traditional diagnostic methods for chronic diseases include a combination of blood tests, imaging studies, and physical examinations. For example, diabetes is typically diagnosed through fasting blood glucose levels, HbA1c tests, or oral glucose tolerance tests. Cardiovascular diseases are often identified through electrocardiograms (ECGs), stress tests, or imaging techniques such as echocardiograms and coronary angiography. COPD diagnosis usually involves spirometry, which measures lung function, and imaging tests like chest X-rays or CT scans.

While these methods are effective for diagnosing diseases in symptomatic patients, they have limitations when it comes to early detection, particularly in asymptomatic individuals. For example, many patients with early-stage diabetes or prediabetes may not exhibit any symptoms and thus may go undiagnosed until the disease has progressed. Similarly, early-stage cardiovascular disease may not be detected until a patient experiences a significant event such as a heart attack. The reliance on symptomatic presentation and periodic screenings can result in missed opportunities for early intervention.

2.3. Advent of Machine Learning in Healthcare

Today, Machine Learning (ML) has played an important role in many fields and sectors and one of the sectors is healthcare. The ability of the ML algorithms could be to identify numerous relationships within the big datasets that are complicated and cannot be identified by the basic forms of statistical methods. This ability is particularly useful in chronic diseases where, often, multiple factors-genes, diet, stress, exposure to chemicals, etc., contribute to the manifestation of the disease, and their contribution is not always additive.

It reveals the fact that the behaviour of patients can be predicted through machine learning models based on the previous record of patients and even the probability of the development of diseases before the development of symptoms. The kind of carrying capacity given by this model enables a healthcare provider to spot high-risk patients and probably modify their disease prognosis with preventive measures or early interventions for the advancement of the disease. For instance, ML algorithms can learn from EHRs, laboratory tests, and even digital bracelets and find out that patients are at risk of developing diseases that are similar to diabetes or cardiovascular disease.

2.4 ML Algorithms in Chronic Disease Detection

There are several successful cases of using ML for the detection of chronic diseases, and all the approaches have their advantages and limitations. Here are some of the most prominent ones:

2.4.1. Support Vector Machines (SVM)

SVMs are well suited for classification problems, in general, and specifically for the identification of new cases of chronic diseases. SVMs operate by identifying the best hyperplane that would be able to segregate the data into various classes. SVMs, in particular have been applied to the classification of patients on risk of chronic diseases inclusive of diabetes and cardiovascular diseases. [2] Followed a similar procedure to mine a similar set of factors from diabetic patients' demographic and clinical data to predict the onset of the disease. Thus, the model yielded high accuracy and proved that SVMs are promising for the diagnosis of chronic diseases. Despite all these, SVMs have disadvantages associated with high computational costs for large datasets and if the data is not linearly separable.

2.4.2. Random Forest (RF)

Random Forest (RF) is an ensemble learning algorithm which, during the training phase, builds a multitude of decision trees and, during the prediction phase, returns the mode of the classes (in classification) or the mean prediction of the individual trees (in regression). Therefore, RF is particularly useful for working with big data with many variables, which makes it appropriate for healthcare data many of which are features. [3] Employed RF to build a model aimed at identifying the factors influencing cardiovascular events in a group of patients based on the data from their EHR. RF model had better prediction accuracy than the conventional risk scores like the Framingham Risk Score, and thus, RF has an opportunity to revolutionize chronic disease prediction. RF models are also beneficial in that they cause fewer cases of overfitting than other models and can also handle missing data.

2.4.3. Neural Networks (NN)

A Neural Network (NN) can be described as a category of the ML algorithm designed to imitate the working of the human brain. They are especially useful in capturing nonlinear patterns in data and, therefore, can be very effective in diseases such as COPD or heart failure that have multiple causes. [4] Employed deep neural networks to capture clinical data for analyzing and predicting the occurrence of COPD. According to statistics that were depicted in the model, there is every possibility that NNs would be useful in the early identification of patients with chronic diseases, even without apparent symptoms. However, NNs pose certain challenges: large datasets and

computational power are needed for their work, and therefore, they can hardly be implemented in a clinic.

3. Methodology

3.1 Data Collection

3.1.1. Dataset Description

The empirical data set contains a broad range of predictor and outcome variables related to patient characteristics, diseases, laboratory test results and hospitalizations. The data set accumulates records for multiple years and which contributes to the recognition of chronic diseases.

- **Patient Demographics**: Historical, current, and potential factors include age, gender, race, and status.
- **Medical History**: Current and past medical conditions, including chronic diseases, other diseases concurrently present, treatments currently being taken and any past hospital admissions.
- **Lab Results**: Meal, workout, stress, sleep, and also records of blood glucose level, cholesterol levels, blood pressure, etc.
- **Hospitalization Records**: Discharge and admission, the purpose of treatment and the overall results.

3.1.2 Data Preprocessing

Data preparations are important mainly with respect to the preparation of the data before it is used to train the model. The preprocessing pipeline consists of the following steps:

- **Data Cleaning**: This may entail the elimination of redundant records, inconsistencies or making amendments to the data set. For example, data sets having exceptional values (e.g. negative age or missing critical variables) are either purified or deleted.
- **Handling Missing Values**: This is not merely a scene in healthcare datasets, as it is expected that some data is missing most of the time. Several things are done, such as using mean, median or mode to estimate the absent values, neglecting the rows with missing values, or using models that do not require the disposal of rows or columns of data, such as Random Forest.
- **Normalization**: In real life, normal distributions are applied to the continuous variables, like lab results, in order to avoid that one variable overpowering the model.
- **Feature Selection**: Methods like RFE, in which feature selection is done, or PCA, in which dimensionality is reduced, are used. The use of this process helps in identifying the relevant features thereby leading to better performing models and models that are easier to interpret.

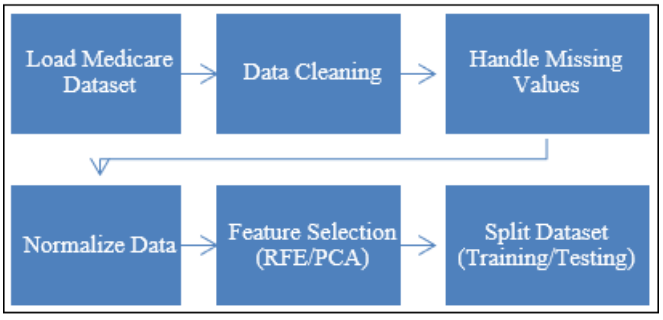


Figure 1: Data Preprocessing Steps

3.2 Selection of Machine Learning Algorithms

In the current world, many services involve machine learning algorithms in their operation. These computing models, referred to as machine learning algorithms, allow computers to learn the patterns in data and then predict using that data [5,6] or make an assessment based on data, all without having to program it. From image and audio recognition and recommendation systems to fraud detection and self-driving cars, including natural language processing, those algorithms are the foundations of today's AI.

3.2.1. Support Vector Machines (SVM)

SVM algorithms are chosen for binary classification tasks; this is important to identify patients in a vulnerable population prone to chronic diseases. The SVM algorithm, in its simplest form, involves the determination of the hyperplane that, when used to classify the two classes, delivers the maximum accuracy. To solve the issue of a non-tertiary relationship between the input variables and the target function, Kernel functions, for instance, the radial basis function (RBF), are used.

3.2.2. Random Forest (RF)

Random forest (RF) is chosen for its stability and capacity to work in high-dimensionality problems with large numbers of features. The working of RF is based on building a number of decision trees and then combining their forecasts to arrive at a final decision. This results in the ensemble, which is very useful in reducing the level of overfitting and increasing the level of generality.

3.2.3 Neural Networks (NN)

Neural Networks (NN), especially Deep Learning models are chosen because of their performance in complex and nonlinear structures within the data set. NNs are a set of layering of neurons which receive information input and predict the output. In this research, the model of choice is a multi-layer perceptron (MLP), a preferred model for classification tasks in health care.

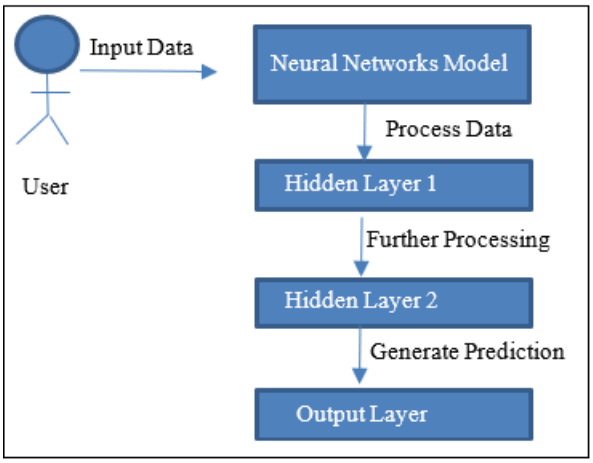


Figure 2: Architecture of a Multi-Layer Perceptron (MLP) Neural Network

3.2.4. Naive Bayes

The Naive Bayes classifier is part of the Bayesian classifier and is built on the principle linked to the Bayes Theorem, which is utilized in the Naive Bayes algorithm family; these

algorithms predicate feature independence. Naively based classifier tends to look at a feature as an independent feature from all the features which are in the same class. When it comes to classification problems, if your classes are more than two but can also be two, Naive Bayes is a good model to employ. When it is assumed that each feature works independently and has an equal contribution of affecting the target class, then the naive Bayes classifier is a probabilistic classifier based on the theory of probability enshrined in the Bayes theorem [7]. The NB classifier assumes there is no relationship between the features, are the features are not responsible for reducing or increasing each other's probability that the sample belongs to a particular class. However, every feature is equally responsible for the probability. It is a decision-pledge theory of fast and efficient results on high-dimensional and large data sets; easy to implement. The NB classifier is robust to noise and is, therefore, useful in real applications.

3.2.5. Logistic Regression (LR)

Logistic regression is a supervised learning process which is majorly applied to classification tasks or assignments. The main aim of this process is to determine the probability of the given instance to belong to the given class. Other classification methods used in it include logistic regression, and interesting to note that this method is actually named so [8]. Thus, using the output obtained from the linear regression function, regression brings out the probability of the provided class using a sigmoid function. This is why regression is effective. Thus, the issue of how the two are related arises because the two concepts need to be reconciled with one another. As opposed to linear regression, which can potentially provide any real value, logistic regression is used to determine whether an instance belongs to a certain class as perceived from the class membership probability assessments.

3.2.6. Random Forest

In both the classification as well as the regression, the Random Forest, which is a supervised learning method is employed. But issues of categorization are at the heart of its performance, this is the kind of problem which its organizational structure is supposed to address and generate. I believe no one will argue that trees are the foundation of a forest, and healthier forests have trees with higher density. On a similar note, the method of the random forest procedure in the data [9,10] samples the construction of decision trees, gets the prediction results from each of the tree types, and opts for the optimal tree with votes. Compared with a single decision tree the ensemble method can be able to average the results and thus be more efficient rather than a single decision tree.

3.2.7. SVM

As for classification and regression in particular, there is a variety in the framework of supervised machine learning algorithms like Support Vector Machine or SVM for short. For classification, however, it really stands out, although there are things we mention about regression as well. HVL The main use of the SVM algorithm is to find the proper hyperplane that can well distinguish the regions of different classes in an (N-dimensional space) [11] The hyperplane wants the greatest possible distance of several classes' nearest points. This is to mean that the number of features decides on the dimensionality of the hyperplane. When there are merely two characteristics for input, then the concept of hyperplane reduces to a line. When three characteristics of the input are incorporated, the hyperplane becomes a two-plane plane.

3.2.8. ExtRa Trees

Extra Trees or Extremely Randomized Trees is another kind of ML that builds a large number of decision trees and uses the result of each of them to determine the final result. However, Random Forest and Extra Trees are almost alike. To ensure that decision trees are sufficiently distinct, Random Forest uses bagging to pick several versions of the training data. On the other hand, Extra Trees builds decision trees on the entire dataset, as shown in the following equation [12,13]. This implies that it may allow, for instance, the feature split, as well as the child node values, to be randomly chosen to make sure that the variation between the decision trees is adequately controlled. Alternatively, in a Random Forest, finding the characteristic splitting value is done by means of a greedy searching algorithm. Random Forest and Extra Trees are very close in terms of these two differences only.

3.2.9. K-Nearest Neighbour (KNN)

However, when it comes to classification and regression, one of the most frequently used methods is the K-Nearest Neighbors (KNN) method. The basic rationale is that labels or values associated with other points of the same dimension are expected to be similar. During the training phase of the KNN algorithm, the whole of the 40 samples of the training dataset [14,15] is saved. To do this, it employs a measure such as the geometric distance so as to measure an input data point's distance to all training samples. After that, the algorithm continues to identify its K nearest neighbors using the distances between the particular input data point and its neighbors. When it comes to methods of classification, the method will assign, as the projected label of the input data point, the most frequent class label among any of the K neighbors. In order to predict an input data point, regression calculates the average or weighted average of the K neighbours' targets. Due to the simplicity and the application of the KNN method in various disciplinary fields, it can be widely used. More crucial for it is the parameter tweaking since its efficiency depends on K and the distance measure used.

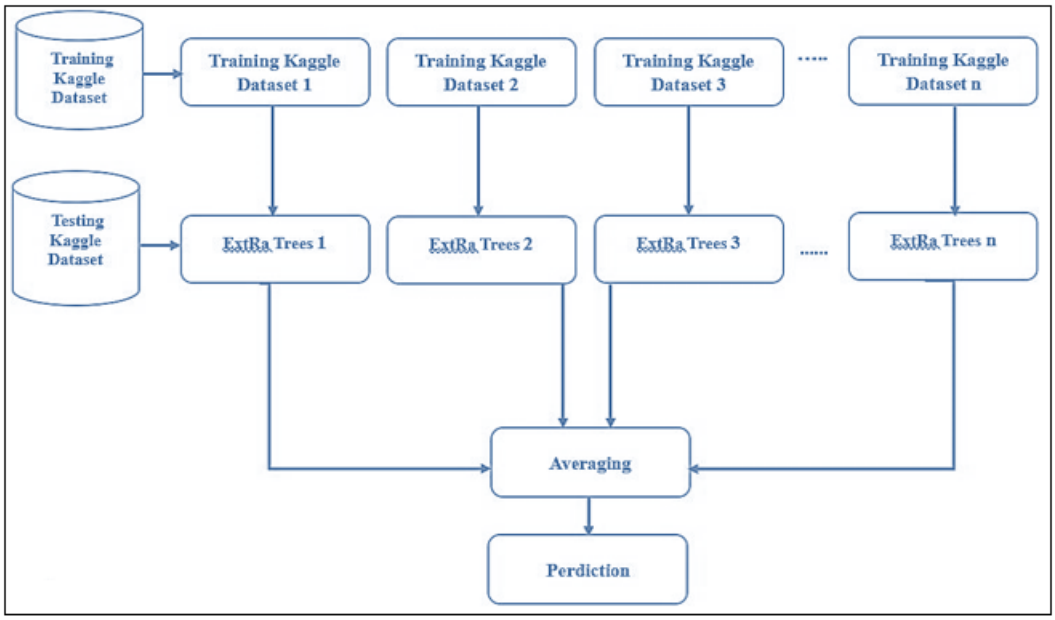
3.3 Recommended Method

Data gathering, cleansing, training model design and assessment are the several steps that may be followed in order to predict diseases with the help of machine learning techniques. In this part, we shall describe the normative process of employing ML for disease prediction. To support the initiation of disease prediction, some data needs to be collected from various data sources, including, but not limited to, electronic health records, patient's genetics, and lifestyle factors. This collected data must then be preprocessed. Data that the system acquires contains noise, missing values as well as outliers, and therefore it is preprocessed.

Figure 3: The Kaggle Data Set

In the selection of the best predictors during the process of illness prediction, the feature selection algorithms are applied. Regarding the type of machine learning that can be employed in disease prediction, the following are possible. The selected approach is trained using supervised learning where the data is preprocessed. For the user to efficiently use the model, hyperparameter optimization is done. I guess in the particular case of human disease prediction with regards to symptoms, what the proposed model is bringing in is a better and more accurate solution. The dataset that was used is depicted in the Kaggle figure**.** 3 ExtRa Trees technique was used to train the models on this dataset. An individual will present himself or herself to the doctor or health care provider with these symptoms. Later in that, we will operate the symptoms through our model. The model will then create the aforementioned potential illness. The novelty of the suggested study is that hyperparameters' tuning increases the efficiency of the ExtRa Trees model. As a result, it is showing more precision. In this study, the author has considered many models using regular data sets for training and evaluation.

3.3.1. ExtRa Trees Algorithm



In the dataset, there are symptoms and the diseases they reflect; for the model training, the ExtRa Trees algorithm is employed. There are several reasons to use the ExtRa Trees [16] algorithm; however, the primary benefit shall be mentioned here, which is the fact that this algorithm can handle datasets containing both continuous and categorical variables. It is especially useful when used in regression as well as in classification tasks. This is where it shines, or rather stands out with unequaled performance when it comes to categorization exercises. As Figure 2 illustrates, ExtRa Trees is a DS which has many advantages for research and development as well as practical implementation, such as improved invocation efficiency, reduced space complexity, and avoidance of add-replace cycles. The first thing which is done is to choose random instances from a given set, or training set. This way, a second-step decision tree will be generated for each training set in the manner described above. Third, the decision tree shall be decided by the average of the trees. The fourth step, as a last step, selects the best forecast that has maximum support. The ExtRa Trees method diagnoses diseases in the following manner: it compares the patient's symptoms with the disease symptoms and the geographical region indicated in the database. The job then proceeds to assess the reliability of the model after the results have been interpreted using the given labels.

```
from sklearn import model_selection
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.tree import ExtraTreeClassifier
from sklearn.metrics import accuracy_score, precision_score, roc_auc_score
# Example dataset (replace with actual data)
# X, y = ...
# Models
models = []
models.append(('SVM', SVC(kernel='rbf')))  # SVM with Radial Basis Function (RBF) kernel
models.append(('RF', RandomForestClassifier()))  # Random Forest
models.append(('NN', MLPClassifier(max_iter=1000)))  # Neural Networks (MLP)
models.append(('ET', ExtraTreeClassifier()))  # Extra Trees
# Evaluation
results = []
names = []
scoring = 'accuracy'
for name, model in models:
    kfold = model_selection.KFold(n_splits=10)
    pipeline = make_pipeline(StandardScaler(), model)
    cv_results = model_selection.cross_val_score(pipeline, X, y, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
      # Fit model and calculate additional metrics
    pipeline.fit(X, y)
    y_pred = pipeline.predict(X)
    accuracy = accuracy_score(y, y_pred)
    precision = precision_score(y, y_pred, average='binary')  # Adjust 'binary' to 'macro' or
'weighted' for multi-class
    roc_auc = roc_auc_score(y, pipeline.decision_function(X))

    print(f"{name} - Accuracy: {accuracy:.2f}, Precision: {precision:.2f}, ROC-AUC:
{roc_auc:.2f}")
```

Figure 4: The Methodology of ExtRa Trees Algorithm

3.4 Model Training and Testing

3.4.1 Training Process

The set is then divided into training and testing sets often in 70:30 ratio bases. [17] The training set on the other hand, is used in the formulation of the models, whereas the testing set, on the other hand, is employed to assess the performance of the models. Cross-validation techniques, especially k-fold cross-validation, are used with a view of tuning parameters as well as to check on cases of over fitting.

- **Cross-Validation**: Dividing the data set into K subsets, the model will be trained on K-1 subsets; at the same time, one subset will be used for validation. This is where the k time's cross-validation process is performed; that is, all the created subsets are used once for validation.
- **Hyperparameter Tuning**: This includes the use of grid search or random search in order to come up with the right hyperparameters for each model. For SVM, for example, one might have to adjust the kernels of function and the parameter of regularization. The number and depth of trees for RF may be modified and this was observed in the experimentation process. For NNs, the number of hidden layers, learning rate, and activation functions are considered practical and optimized.

3.4.2 Testing and Validation

The trained models are then tested on the test set in order to measure the quality of the models. [18] The following metrics are calculated:

- **Accuracy**: The ratio of the number of test instances that have been classified correctly out of the total number of test instances.
- **Sensitivity (Recall)**: The integrity of the propositions used to determine positive individuals, that is, the sufferers of a certain chronic disease.
- **Specificity**: The overall health of individuals misclassified by the model, for instance, negative cases such as healthy people.
- **ROC-AUC**: The Area under the Receiver Operating Characteristic curve gives a general idea of the performance of the model over a range of classification thresholds.

3.5.1 Accuracy

Accuracy is a Common evaluation metric that estimates the total variation of the model's success rate out of the total entities and defines it as the ratio of the correct observation out of the total observation. [19-21] Accuracy, though helpful, is inadequate on its own, especially when applied to cases where the number of negative observations is considerably higher than that of positive ones.

3.5.2 Sensitivity and Specificity

- **Sensitivity (Recall)**: Specificity measures the proportion of actual negatives which the model classifies as such. High sensitivity is essential, especially for patient health risk applications, so that affected patients are appropriately marked for further diagnostics or treatment.
- **Specificity**: Specificity measures the ratio of true negatives that is, how many individuals who are level one classification are truly negative solutions. In the setting of chronic disease identification, specificity then serves to exclude individuals who do not have the disease of interest to reduce their likelihood of being intervened on.

3.5.3 ROC-AUC

The ROC-AUC (Receiver Operating Characteristic - Area Under Curve) is a metric that provides a comprehensive view of a model's performance across various classification thresholds. A model with an AUC close to 1 is considered highly effective, while an AUC close to 0.5 indicates a model with no discriminative ability.

3.6 Implementation in Clinical Settings

3.6.1 Feasibility Analysis

Before deploying these ML models in clinical environments, a preliminary study of the practical realization of the application is carried out. This also involves the assessment of the computational utility of these models, including possible real-time computation of these models, their accommodation into existing clinical processes, and clinician uptake of these models.

- **Computational Requirements**: Some of the models may take time and computational resources especially when they have to be used for a large database such as NNs. The use of cloud services or obtaining high-performance computing systems on a company's site may be required.
- **User Acceptance**: Some central issues are health care provider training to comprehend and rely on ML predictions. This may require designing user interfaces for models which would display them and their confidence levels in an easily understandable format.

3.6.2 EHR Connectivity

Neither of the modeling methods can be solely useful in addressing the challenges in healthcare; consequently, integrating ML models into EHR systems is important to offer timely decision support to clinicians. This integration enables the ML models to review patient data as they fill the system and give suggestions or risks at that particular moment.

- **Real-Time Decision Support**: The patients with a higher risk of getting chronic diseases and the ML models can raise an alarm that can make clinicians order other tests or even take preventive measures.
- **Data Flow**: The integration works in the sense that the data must be flowing between the EHR system and the ML model to avoid any delay in the analysis of the patient data.

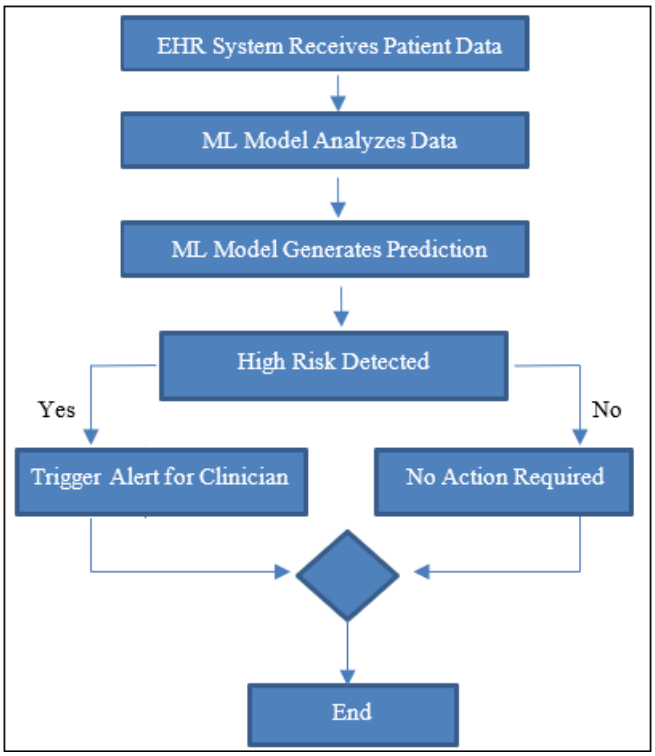


Figure 5: Integration of ML Models with EHR Systems

4. Results and Discussion

The findings of this study are shown in this section in relation to the efficacy of three machine learning techniques, namely SVM, RF, and NN to identify chronic diseases among the Medicare population. Some of the major diseases for which there is a need to carry out comparative analysis are diabetes, cardiovascular disease, COPD and heart disease. The objective was to determine the model with the best predictive performance in terms of chronic disease diagnosis and generalization, but also with consideration of its computational requirements and its sensitivity to different sources of bias.

4.1 Comparative Analysis

4.1.1 Model Comparison

- **SVM**: Showed the highest degree of sensitivity and specificity in the diagnosis of diabetes and heart disease. It is particularly appropriate for high-dimensional spaces and, however, is rather time-consuming.
- **RF**: Reported the best results with data size and features that are used in terms of both performance and computational complexity. It was especially effective in the diagnosis of cardiovascular diseases.
- **NN**: Appeared to be most accurate for identifying higher-order, nonlinear relations, especially those concerning COPD prediction. However, it called for more data and computational power to achieve higher accuracy.

4.1.2 Discussion of Results

This research implies that by adopting machine learning, it is possible to enhance the early diagnosis of chronic diseases among Medicare individuals. However, this should depend on the application and characteristics of the data that is to be used in the model. For example, where high levels of

accuracy are needed, such as in text classification, SVM could be used, while for tasks that need moderate accuracy, and this is in addition to moderate computational resources, RF would suffice. NN is particularly useful when used to detect diseases that present certain patterns; however, it uses more computing power.

4.2 Challenges

4.2.1 Algorithmic Bias

There is an issue of generalizability because of the algorithm risk, especially if the training data does not cover all Medicare subgroups. For instance, if some demographic categories are excluded, then the model may not be good at predicting diseases in such categories. However, to avoid such a bias, it is inevitable to have a diverse and representative data set.

4.3 Future Research Directions

It is proposed that in future research the data sample should be expanded to include younger populations also, which would make the development of the models more generalizable. Also, it may be useful to investigate appealing other approaches of the machine learning methodologies, including gradient boosting or deep learning models, for the enhancement of predictive discrimination. Other issues, such as algorithm bias and data quality, will also have to be focused on to make sure that such models can readily be applied in clinical practice.

5. Conclusion

From this research, it is shown that machine learning techniques such as SVM, RF and NN can play a vital role in improving the case-detection of chronic diseases in the Medicare population in the early stages. Each model brings unique strengths to the table: SVM is best suited in terms of accuracy, especially with diagnoses such as diabetes and heart disease; RF is best used with large datasets complex in features which it balances performance and time; and NN in capturing nonlinear complex patterns in conditions such as COPD. Nevertheless, there are several acknowledged problems, including the demand for superior and generalizable data, the possibilities of an algorithm's prejudice, and concerns about model explainability. This paper explores several challenges as follows, which must be solved to optimize the adoption of ML models in tracheostomy care. For the healthcare system, the prospects of deep ML models as a new approach to chronic disease management through early detection and timely interventions can have a massive impact on the results and costs. With healthcare being a wide and dynamic field in the present generation, the employment of machine learning in enhancing resource utilization and patient care delivery is most likely to be a critical development in the enhancement of efficiency, effectiveness, and tailored healthcare solutions in the future.

References

[1] Chronic obstructive pulmonary disease (COPD), 2023. Online. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(COPD)

[2] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-diabetes. BMC Medical Informatics and Decision Making, 10(1), 16.

[3] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can Machine Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? PLOS ONE, 12(4), e0174944.

[4] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Nature, 542(7639), 115-118.

[5] Y. Perwej, Md. Husamuddin, Fokrul Alom Mazarbhuiya,"An Extensive Investigate the MapReduce Technology", International Journal of Computer Sciences and Engineering (IJCSE), Volume-5, Issue-10, Page No. 218-225, 2017, DOI: 10.26438/ijcse/v5i10.218225

[6] N.Akhtar, Devendera Agarwal, "An Efficient Mining for Recommendation System for Academics", International Journal of Recent Technology and Engineering (IJRTE), ISSN 2277-3878 (online), SCOPUS, Volume-8, Issue- 5, Pages 1619-1626, 2020, DOI: 10.35940/ijrte.E5924.018520

[7] C. Zhenhai and Liu. Wei, "Logistic Regression Model and Its Application", Journal of Yanbian University (Natural Science Edition), vol. 38, no. 01, pp. 28-32, 2012

[8] Firoj Parwej, Nikhat Akhtar, Y. Perwej, "A Close-Up View About Spark in Big Data Jurisdiction", International Journal of Engineering Research and Application (IJERA), ISSN: 2248-9622, Volume 8, Issue 1, (Part -I1), Pages 26-41, January 2018, DOI: 10.9790/9622- 0801022641

[9] M. Liu, X. Xu, Y. Tao and X. Wang, "An improved random forest method based on RELIEFF for medical diagnosis", 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 1, pp. 44-49, 2017

[10] R. Cuingnet, C. Rosso, M. Chupin, S. Lehéricy, D., H. Benali, et al., "Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome", Medical Image Ana., vol. 15, no. 5, pp. 729-737, 2011

[11] M. R. Camana Acosta, S. Ahmed, C. E. Garcia and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks", IEEE Access, vol. 8, pp. 19921-19933, 2020

[12] Y. Perwej, "Unsupervised Feature Learning for Text Pattern Analysis with Emotional Data Collection: A Novel System for Big Data Analytics", IEEE International Conference on Advanced Computing Technologies & Applications (ICACTA'22), SCOPUS, IEEE No: #54488 ISBN No Xplore: 978-1-6654-9515-8, Coimbatore, India, 2022, DOI: 10.1109/ICACTA54488.2022.9753501

[13] Shobhit Kumar Ravi, Shivam Chaturvedi, Dr. Neeta Rastogi, N. Akhtar, Y. Perwej, "A Framework for Voting Behavior Prediction Using Spatial Data", for

published in the International Journal of Innovative Research in Computer Science & Technology (IJIRCST), ISSN: 2347-5552, Volume 10, Issue 2, Pages 19- 28, 2022, DOI: 10.55524/ijircst.2022.10.2.4

[14] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors", IEEE Trans. neural networks Learn. Syst., vol. 29, no. 5, pp. 1774-1785, 2017

[15] Y. Perwej, Md. Husamuddin, Dr. Majzoob K.Omer, Bedine Kerim, "A Comprehend the Apache Flink in Big Data Environments", IOSR Journal of Computer Engineering (IOSR-JCE), e- ISSN: 2278-0661, P-ISSN: 2278-8727, USA, Volume 20, Issue 1, Ver. IV, Pages 48-58, Feb. 2018, DOI: 10.9790/0661-2001044858

[16] Y. Perwej, Dr. S.A. H., Firoj Parwej, Nikhat Akhtar, "A Posteriori Perusal of Mobile Computing", International Journal of Computer Applications Technology and Research (IJCATR), which is published by ATS (Association of Technology and Science), India, ISSN 2319–8656 (Online), Volume 3, Issue 9, Pages 569 - 578, 2014, DOI: 10.7753/IJCATR0309.1008

[17] Centers for Medicare & Medicaid Services (CMS). Chronic Conditions among Medicare Beneficiaries. Online.https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/Downloads/2012Chartbook.pdf]

[18] Centers for Disease Control and Prevention (CDC). National Diabetes Statistics Report, 2020. Online. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf]

[19] National Institutes of Health (NIH). COPD: A Leading Cause of Death in the United States. Online. https://www.nhlbi.nih.gov/health-topics/copd]

[20] American Diabetes Association. Standards of Medical Care in Diabetes—2020. Diabetes Care, 2020; 43(Suppl. 1): S1-S212.

[21] National Heart, Lung, and Blood Institute (NHLBI). What Is Coronary Angiography? Online. https://www.nhlbi.nih.gov/health-topics/coronary-angiography]

[22] Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management, and Prevention of COPD, 2020. Online. https://goldcopd.org/gold-reports/

[23] American Diabetes Association. Diabetes and Cardiovascular Disease. Diabetes Care, 2018; 41(Suppl. 1): S86-S104.