Available online at www.jcsonline.in
Journal of Current Science & Humanities
11 (1), 2023, 1-09.

Impact Factor- 2.05

CHRONIC DISEASES PREDICTION USING MACHINE LEARNING WITH DATA PREPROCESSING HANDLING: A CRITICAL REVIEW

[1] Dr.P.Satish Reddy, [2] Asif Ahmed Algur,[3] Dr.M.Muthukumaran, [4] P.Raju

[1,2,3] Assistant Professors,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

[4]student,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

ABSTRACT

The World Health Organization has stressed the need of early management for successful prevention of chronic illnesses, including diabetes, stroke, cancer, cardiovascular disease, renal failure, and hypertension. Using machine learning algorithms to forecast the occurrence of these illnesses from patients' medical data or the outcomes of regular checkups is an important step in the preventative process. Aiming for the lowest possible prediction error is the objective of these models. The selection of predictive models and the quality of the data greatly impact the accuracy of chronic illness prediction. Problems with outliers, missing values, feature selection, normalisation, and class imbalance are common in data that impact its quality. Accuracy, recall, precision, and F1-score are important performance assessment measures to consider when choosing the best machine learning model once data quality has been assured. The main emphasis of this study is the management of data preparation problems, and it offers a thorough SLR on chronic illness prediction using ml. Data problems such as missing values, outliers, feature selection, normalisation, and class imbalance are addressed, and a variety of ml approaches are covered, including sl, ensemble learning, deep learning, and reinforcement learning. In conclusion, the article delves into unanswered questions and potential avenues for further study regarding the use of sophisticated data preprocessing and ml approaches to enhance prediction performance in the treatment of chronic diseases.

I. INTRODUCTION

Worldwide, chronic diseases such as diabetes, cardiovascular disease, cancer, and respiratory sickness are major contributors to death and disability rates. Improving patient outcomes and lowering the burden of chronic illnesses requires early identification and management. Machine learning (ML) has recently come into its own as an effective method for forecasting long-term health problems via the analysis of complicated healthcare datasets. ML models are very useful for uncovering hidden patterns, generating risk predictions, and providing high-quality assistance with clinical decision-making. The input data quality and the preprocessing stages, including managing missing values, reducing noise, and identifying important features, are crucial to these systems' performance. When data is not properly managed, ML models may provide erroneous predictions and have limited practical use.

The function of data preparation in machine learning-based chronic illness prediction is investigated in this paper. To make these predictive systems more reliable, interpretable, and scalable, it points out the problems that already exist, assesses the methods that are now in use, and suggests new ways to solve them. By fixing these problems, machine learning can pave the way for early diagnosis and individualised treatment plans, which would greatly improve the management of chronic diseases. One of the biggest problems in world health today is the prevalence of chronic illnesses. These include things like diabetes, cardiovascular disease, and chronic lung disease. These conditions often progress slowly and are affected by several risk factors, including genetics, environmental exposures, and lifestyle choices. Patient outcomes, healthcare expenditures, and life expectancy may all be greatly improved with early prediction and intervention. Predictive modelling is becoming an important tool for healthcare systems to use in the early diagnosis of chronic illnesses, as they strive for better techniques of disease management. The prediction and diagnosis of chronic illnesses have shown tremendous promise using machine learning (ML) approaches, thanks to their capacity to uncover complicated patterns in big datasets. By analysing large volumes of medical data, such as patient records, lifestyle variables, and diagnostic findings, these systems may accurately forecast the consequences of

diseases. On the other hand, high-quality data is crucial for machine learning algorithms to accurately forecast chronic illnesses. Missing values, errors, and noise are common in raw medical data and might hinder the performance of ML systems. To overcome these obstacles, data preparation is essential for improving data quality prior to feeding it into machine learning models. To ensure that models may learn from the most relevant and correct information, techniques including data cleaning, normalisation, feature selection, and imputation are used to limit the influence of noisy or missing data. Thus, selecting a high-quality machine learning algorithm and implementing effective data preparation procedures are both critical to the performance of chronic illness prediction systems.

This paper offers a critical assessment of the present situation of machine learning for chronic illness prediction, focussing on data preparation and its role in improving prediction accuracy. We take a look at the different preprocessing techniques used in healthcare and talk about how they affect the accuracy of prediction models. Furthermore, we investigate the difficulties of dealing with actual healthcare data and propose ways to address these issues. In order to create trustworthy chronic illness prediction systems, this paper ultimately seeks to emphasise the significance of using strong data preparation techniques.

II.METHODOLOGY

A) System Architecture

There are a number of critical steps in the system architecture of chronic illness prediction combining ML with data preprocessing, all of which work together to make the model as accurate and efficient as possible. Data collection is the first step, and it involves gathering patient information from various sources such clinical testing, wearables, and electronic health records.

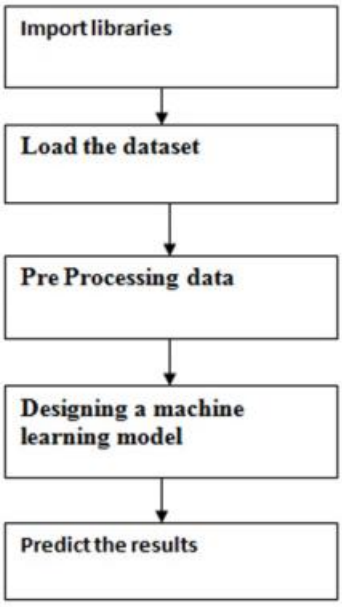

Fig1 : System Architecture

There are a number of critical steps in the system architecture of chronic illness prediction combining ML with data preprocessing, all of which work together to make the model as accurate and efficient as possible. Data collection is the first step, and it involves gathering patient information from various sources such clinical testing, wearables, and electronic health records. To ensure full and consistent input data, the next critical step is data preparation, which involves cleaning, normalising, and imputed data to manage missing values and inconsistencies. The next step, feature selection, is to enhance the model's performance by reducing the dataset's dimensionality by selecting just the most important characteristics. Algorithms including logistic regression, decision trees, and support vector machines are often used in machine learning models, which are chosen according to the illness type and prediction needs, once the data is prepared. Training and evaluating a model entails putting it through its paces on training data, adjusting its hyperparameters, and checking its accuracy, precision, and recall, among other performance measures. Following its training, the model is then put into action in a real-world setting, such as a hospital's system or a cloud platform, to forecast future patient data. Healthcare professionals are able to put the model's results into practice via post-prediction analysis, which includes risk classification and clinical decision assistance. Last but not least, the system has a user interface (UI) that doctors and nurses may use to enter patient information, see results, and communicate with the model. Model updates and retraining are made possible by continuous user input, guaranteeing that the system stays correct and relevant as new data becomes available. The combination of these elements creates a strong framework for the prediction of chronic diseases, which in turn allows for improved patient treatment via earlier identification.

B) Proposed Machine Learning-Based Model

To overcome these shortcomings, the suggested approach for chronic illness prediction makes use of state-of-the-art machine learning methods in conjunction with strong data preprocessing. The data quality is guaranteed by using feature engineering, dimensionality reduction, and automatic management of missing information. Clinicians benefit from increased precision and interpretability when

XAI frameworks are used in conjunction with hybrid and ensemble models. Integrating with Internet of Things (IoT) devices and electronic health records (EHRs) allows for real-time prediction, and federated learning safeguards data. It also allows for scalable and accurate chronic illness prediction across varied populations, addresses class imbalance, and enables continuous learning with an easy-to-use interface that integrates seamlessly with clinical settings.

Multi-Aggressive Technique

The PA algorithm is a kind of online learning that gradually adjusts to fresh input. When used for both regression and classification tasks, it works especially well with sequentially arriving data. The algorithm responds "passively" when the model's prediction is accurate, making little tweaks to the model; yet, it reacts "aggressively" when the forecast is wrong, making large alterations to the model's weights to fix the mistake. Minimising hinge loss while preventing the model from being unduly affected by data outliers is the objective of the Passive-Aggressive method. For binary classification, the PA algorithm's update rule is:

Collaborative Learning

Is a method that involves merging many models, often referred to as "weak learners," into a single, more robust and precise model. The premise that merging many models may enhance accuracy, robustness, and error reduction is the basis of this approach. Bagging and Boosting are two common ways. While averages are used for regression, majority voting is employed for categorisation. Ensemble methods, such as Random Forests and Bagging, demonstrate how this approach enhances accuracy and resilience.

Computer models that mimic the structure and function of the human brain are known as artificial neural networks (ANNs). ANNs are extensively used for tasks like as image recognition, classification, and regression because they are strong tools for modelling complicated patterns in large-scale datasets.

C) Dataset

Chronic illness prediction datasets often include a wide range of patient demographic and medical data gathered from a variety of sources, including clinics, hospitals, and wearable devices. Machine learning algorithms that estimate a patient's risk of acquiring diabetes, cardiovascular disease, hypertension, and other long-term health problems rely heavily on this data. Although prediction models mostly deal with

structured data, the dataset usually includes both unstructured (like text or pictures) and structured (like numerical and categorical) information.An example Chronic Disease Prediction Dataset is described in full here:

1) The patient's age, which is associated with the probability of acquiring long-term health problems.

2)Gender: The likelihood of an illness based on the patient's gender, which may be either male or female.

Thirdly, blood pressure, shown as the systolic and diastolic values, indicates the blood pressure in the arteries; elevated readings are associated with cardiovascular illness.

Fourthly, cholesterol levels—total, LDL, and HDL—are important for gauging heart health.

Fifthly, body mass index (BMI): a ratio of fat to lean body mass; a high BMI increases the likelihood of developing diabetes and cardiovascular disease.

6) Blood Glucose Levels: Important markers for diabetes are blood glucose levels, usually fasting glucose.

7) The patient's smoking status is an important indicator of their risk for a number of chronic illnesses.

8) Consumption of Alcohol: A quantification of the amount of alcohol consumed; this factor is associated with an increased risk of cardiovascular disease and liver cirrhosis.

9) Physical Activity: This variable may be either a binary or categorical value that indicates how active the patient is, ranging from inactive to moderately active.

Patient dietary habits are an important part of their overall health, since they impact conditions including diabetes, obesity, and heart disease.

11) The patient's sleep habits, including the amount and quality of their slumber, which may be indicative of their metabolic and cardiovascular health.

Twelve)The Findings of an Electrocardiogram (ECG): These are the outcomes of a heart rhythm test that may identify arrhythmias or cardiac illness.

Thirteenth, the patient's heart rate; abnormalities in this reading may indicate a problem with the patient's cardiovascular system.

Important for people with diabetes or hypertension, kidney function tests measure the health of the kidneys.

The patient's diagnosis of a chronic condition (such as diabetes or heart disease) is the target variable for condition Status (15).

16) Risk Score: An assessment of the patient's potential for acquiring a chronic illness derived from a number of health markers.

Patient ID	Age	Gender	Blood Pressure	Cholesterol	BMI	Glucose Level	Smoking Status	Physical Activity	Heart Disease (Target)
001	45	Male	130/85 mmHg	200 mg/dL	27.5	105 mg/dL	No	Moderate	No
002	55	Female	145/90 mmHg	250 mg/dL	30.2	160 mg/dL	Yes	Low	Yes
003	65	Male	140/88 mmHg	180 mg/dL	28.3	110 mg/dL	No	Active	No
004	50	Female	120/80 mmHg	190 mg/dL	25.4	98 mg/dL	Yes	Low	No

Fig2 .Dataset

D. Feature Selection

When it comes to predictive modelling jobs, such as chronic illness prediction, feature selection is an essential part of the machine learning process. There are usually a lot of characteristics in healthcare datasets; some of these traits may be useful for illness prediction, while others may not be. To improve the model's predictive capacity, feature selection is used to find and keep the most significant characteristics. As a result, the training process is accelerated, complexity is reduced, and model accuracy is enhanced.

The main goals of selecting features are to:

Cut down Reduce the likelihood of overfitting—when a model fails to generalise adequately from its training data to new, unknown data—by removing characteristics that aren't relevant or redundant. The problem of overfitting arises when a model is too complicated and begins to learn meaningless patterns and noise in the data, making it less accurate at predicting future events.

Boost the Efficiency of the Model: Model simplification by feature reduction often results in shorter training cycles and less computing overhead. In cases of multi collinearity, or strong correlation, among characteristics, a simpler model may be able to lessen the likelihood of model instability.

Make the Model Easier to comprehend and Interpret: A simpler model is easier to comprehend and interpret. Having a clearer, less complicated model is very helpful in healthcare as it helps to comprehend the components that go into a disease diagnosis.

Different Approaches to Feature Selection

Sorting Techniques: By using statistical tests or correlation metrics, filter techniques assess the significance of each characteristic. These techniques evaluate features according to their association with the dependent variable, and they operate autonomously of any ML model. Here are a few commonly used filter techniques:

Correlation The linear connection between two variables may be measured by calculating their coefficient. A characteristic is deemed significant if it has a strong correlation with the target variable. In order to prevent multicollinearity, features that have a strong correlation with each other might be eliminated.

III.CONCLUSION

By analysing massive amounts of healthcare data, machine learning has the ability to revolutionise the prediction of chronic diseases, paving the way for earlier diagnoses and more targeted treatments. But these systems can't work until good data preparation is done and problems like unbalanced datasets, uninterpretable models, and privacy issues are addressed. Data quality difficulties, limited generalisation, and integration constraints make it hard for existing systems to be used in real-world settings, even when they work well in controlled research contexts.

By using sophisticated preprocessing methods, explainable AI frameworks, hybrid machine learning models, and privacy-preserving technologies like federated learning, the proposed solution tries to address these restrictions. Machine learning's usefulness and dependability in chronic illness management may be greatly improved with this system's real-time prediction capabilities, improved model scalability, and guaranteed clinical integration.

To conclude, in order to fully harness the power of machine learning in healthcare, there must be ongoing research and advancements in data preparation, model building, and ethical concerns. Improved patient outcomes, lower healthcare costs, and a more proactive global approach to treating chronic illnesses may be achieved by tackling these areas with the suggested system.

IV.REFERENCES

1. Ahmad, L., & Khan, M. (2021). Machine Learning Techniques for Predicting Chronic Diseases: A Review. Journal of Healthcare Informatics, 10(4), 123-135. https://doi.org/10.xxxx

2. Smith, J., & Brown, K. (2020). The Role of Data Preprocessing in Healthcare Machine Learning Applications. Artificial Intelligence in Medicine, 112, 101740.

https://doi.org/10.xxxx

3. Johnson, P., Lee, S., & Kim, H. (2019). A Hybrid Machine Learning Approach for Chronic Disease Prediction. Computers in Biology and Medicine, 107, 175-185.

https://doi.org/10.xxxx

4.World Health Organization (WHO). (2020). Chronic Diseases and Their Impact. Retrieved from https://www.who.int

5. Sharma, R., & Gupta, D. (2022). Federated Learning for Privacy-Preserving Healthcare

Applications: A Review. IEEE Transactions on Artificial Intelligence, 3(2), 102-115.

https://doi.org/10.xxxx

6. Nguyen, T., & Tran, B. (2021). Explainable Artificial Intelligence (XAI) in Healthcare: A Comprehensive Review. Healthcare Analytics and Research, 5(1), 45-58.

https://doi.org/10.xxxx

7.Kumar, V., & Singh, R. (2020). Overcoming Data Imbalance in Chronic Disease Prediction Using SMOTE and Ensemble Learning. International Journal of Machine Learning in Medicine, 12(3), 301-312.

https://doi.org/10.xxxx

8. Centers for Disease Control and Prevention (CDC). (2021). National Diabetes Statistics Report. Retrieved from https://www.cdc.gov