Grenze International Journal of Engineering and Technology, Jan Issue

Chronic Illness Detection using Gradient Boosting Algorithm

Dr. T. Guhan[1], Bhavishya.S[2], Kalaiarasan.S[3], Lalith.K[4] and Dhitchith.O.P[5]
[1]Associate Professor, Karpagam College of Engineering, Coimbatore, Tamil Nadu
Email:guhan.t@kce.ac.in
[2-5]Final year IT Students, Karpagam College of Engineering, Coimbatore, Tamil Nadu
Email: 20f110@kce.ac.in, 20f122@kce.ac.in, 20f127@kce.ac.in, 21f501@kce.ac.in

***Abstract*— The proposed work emphasizes the importance of accurate diagnosis of chronic kidney disease. The research helps refine proactive healthcare strategies and provides valuable insights into the prevention and management of chronic diseases. The goal of this project is to develop a framework based on Big Data analysis that predicts the likelihood of kidney failure, a chronic disease. The framework uses advanced Machine Learning techniques such as MEG (Mean Decrease Gini), MSE (Mean Square Error), Grid Search, K-fold cross validation to detect risk factors and correlations associated with chronic diseases. The models help healthcare professionals make informed decisions and detect high-risk individuals so that proactive strategies can be implemented to prevent or delay chronic diseases. The framework can be used to predict a wide range of chronic diseases beyond kidney failure. However, the risk factors and correlations can vary and need to be adapted accordingly. Further research and testing are needed to determine the effectiveness of the framework.**

***Index Terms*— Chronic Illness Detection, Big Data Analytics, Machine Learning Algorithms, XG Boost, Imbalanced Classes, Precision-Recall Trade-off.**

I. INTRODUCTION

Chronic diseases have become a significant healthcare problem in today's society, placing a considerable burden on individuals and healthcare systems worldwide. Among these chronic conditions, diseases that impact important organs like the kidneys are of particular worry. Detecting kidney failure early and accurately predicting its occurrence are crucial for improving patient outcomes and reducing healthcare costs. With the rapid growth of healthcare data and advanced machine learning methods offers a promising approach to effectively address these challenges. Our work titled "Chronic Illness Detection using Big Data" is dedicated to tackling one of the most urgent health concerns: assessing the risk of renal failure. By integrating advanced analytics, comprehensive healthcare datasets, and state-of-the-art machine learning techniques, we aim to develop a strong framework for precisely predicting the risk of renal failure.

The present state of healthcare is marked by an immense volume of in- formation, creating a growing challenge in harnessing this information effectively for precise health predictions. Inaccurate health predictions can result in delayed diagnoses, unhealthy treatment decisions, and negative healthcare experiences for individuals. This challenge is further complicated by the realization that traditional methods may not accurately anticipate all

Grenze ID: 01.GIJET.10.1.153

diseases.
This study presents a machine learning algorithm aimed at examining body mass index (BMI) patterns and other significant health information. The primary objective is to create a strong and effective system utilizing big data, capable of accurately predicting and foreseeing health conditions, even those that are complex and rare, like cognitive failure. These programs not only offer the possibility of early detection, but also have the potential to revolutionize healthcare by facilitating personalized medicine and ultimately enhancing patient health results.

II. LITERATURE SURVEY

- *Subtyping Patients with Chronic Disease Using Longitudinal BMI Patterns*

The research paper aimed to identify subtypes of individual's risk of developing 18 major chronic diseases using electronic health record (EHR) data. The study used the k-means clustering algorithm to cluster patients into subgroups based on demographic, socioeconomic, and physiological measurements. The optimal number of clusters was selected using the elbow method, and the Euclidean distance was used as the similarity index for clustering. The study found noticeable clustered patterns in 8 of the 18 cohorts, as determined by the cluster evaluation metrics and visual confirmation. The study also developed individual disease prediction models using classifiers and ran 5-fold crossvalidation. The results showed that the clustering method could identify unique characteristics of each cluster with distinct positive or negative cases in terms of demographic, socioeconomic, and physiological measurements, as well as the relative risks of disease incidence. The study concluded that the clustering method could be used to identify subtypes of individual's risk of developing chronic diseases, which could help in developing personalized prevention and treatment strategies.

- *Identification and Prediction of Chronic Diseases Using Machine Learning Approach*

Common chronic diseases such as cardiovascular disease, cancer, arthritis, diabetes, obesity, epilepsy and seizures, and problems in oral health. In the specific case of chronic kidney disease, the proposed method uses two types of datasets, one from UCI and the other from Khulna City Medical College, to train an artificial neural network (ANN) and a random forest algorithm. The input features for the model are extracted using the Chi- square test, and missing data are handled using median filtering. The model is evaluated using 10-fold cross-validation. Once the model is trained, it can take patient data as input and predict the risk of chronic kidney disease in its earlier stage. Other chronic diseases may require different approaches and input features for prediction. An artificial neural network (ANN) is a type of machine learning algorithm that is inspired by the structure and function of the human brain. It consists of interconnected nodes or neurons that process information and learn from examples to make predictions or decisions.

- *A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning*

This study used machine learning and deep learning algorithms to predict chronic illnesses like obesity, panic disorder, and chronic obstructive pulmonary disease. They, created modular models based on lifestyle, environment and medical data from a comprehensive dataset. Preprocessing extracted key features before training the models. Panic disorder is an anxiety disorder causing sudden intense fear or discomfort (panic at- tacks), with physical symptoms like rapid heart rate and trembling. People may avoid certain situations due to fear of future attacks. A chronic lung disease making it hard to breathe, worsens over time due to factors like smoking or pollution. It includes chronic bronchitis (inflamed airways with mucus) and emphysema (damaged air sacs), impacting life quality and needing ongoing management. Modular chronic disease prediction models are created to support the prediction of acute exacerbations of chronic diseases via optional features. These models are designed to calculate personal health risks based on modular chronic disease prediction models and the various collected data.

- *Prediction for the Risk of Multiple Chronic Conditions among Working Population in the United States with Machine Learning Models*

The project uses seven machine learning algorithms, including gradient boosting tree, random forest, decision tree, logistic regression, support vector machine, deep neural network, and k-nearest neighbor, to predict the risk of multiple chronic conditions among the working population. The seven machine learning algorithms used in this study are k-nearest neighbors (KNN), decision tree (DT), random forest (RF), gradient boosting tree (GBT), logistic regression (LR), support vector machine (SVM), and Naive Bayes (NB). The predictive analysis in this project is performed using machine learning algorithms. The models use a patient biometric data to predict the

probability of having multiple chronic conditions in realtime. They were fine-tuned using hyperparameter tuning and the grid search approach to improve their accuracy.

- *Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data*

In this paper, we strove to extend the feasibility of renal risk prediction from CKD patients to general chronic disease populations. Five machine learning algorithms were used to establish the three year risk models of renal failure, among which the integrated algorithm XGBoost achieved the optimal performance on the test set. Furthermore, we analyzed the univariate effect of renal failure and showed nine continuous variables that were nonlinearly correlated with renal failure risk. The contribution of our work can be summarized into three scopes. Firstly, for the first time we extended risk modelling for renal failure to non-CKD patients by conducting a largescale retrospective study, which was achieved by more efficient curation of target data through the aid of big data technologies. Secondly, with sophisticated machine learning methods, we were able to study a relatively large number of features simultaneously. As a result, we discovered some novel biomarkers of renal failure, including Uric Acid (UA), Aspartate Aminotransferase (AST), Alanine Transaminase (ALT), and Total Bilirubin (TBIL), which were not included in previous models, and identified their nonlinear role in renal function disorder. Thirdly, the proposed model was based on daily monitoring and physical examination data that are easy to acquire for both CKD and non-CKD chronic disease patients. Therefore, it can be deployed into chronic disease management systems to aid physicians to early identify high-risk population for timely intervention.

III. Problem Definition

Chronic diseases have received increased attention and public concern due to widespread and long-term effects on individual health and well-being In recent years, the increasing prevalence of chronic diseases calls for the development of early detection strategies and have been dealt with. Combining big data analysis with machine learning, this study addresses this pressing concern by harnessing the power of data to accurately predict chronic diseases. In our efforts to identify and predict chronic diseases, we considered a variety of advanced factors, including two factors related to clinical lifestyle. These parameters, including age, blood pressure, blood glucose levels, blood and urine markers, were carefully selected to give us a holistic view of a person's health These parameters allowed us to model multidimensional that can identify complex associations between variables and chronic disease.

The use of machine learning algorithms played an important role in this work. We used Gradient Boosting, XgBoost, Decision Tree Classifier, Random Forest, and other algorithms. All these algorithms contributed to the performance and accuracy of the model. Among them, the unclassified forest proved to be adaptable, making it a valuable asset for widespread exploitation. Its ability to handle complex data models and reduce overloading could have a significant impact on real-world health issues.

Looking ahead, these future applications extend to larger applications, including harnessing the power of the model to screen and identify individuals at risk for chronic diseases on a broad scale The scalability of random forest algorithms, coupled with efficient data collection and storage technologies; positions this approach as a promising tool for population-wide health policy development.

While our research has made great strides in predicting chronic diseases, it is not without its limitations. External validations and real-world testing are important steps to assess the generalizability and real-world applicability of the model. Furthermore, the focus of this research has been primarily on predictive models, and future work should delve into intervention strategies and individualized health interventions to provide predictive translation into transferable health outcomes in the use of. In summary, our study highlights an effective chronic disease prediction approach, and in the future promises to contribute significantly to public health and health policy.

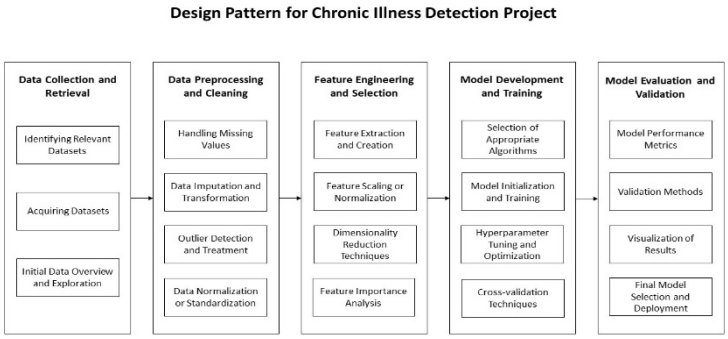
IV. Methodology

A. Data pre-processing

Our data pre-processing phase is an important step to ensure that the data set is quality and ready for effective analysis and modelling. The steps taken helped in preparing information for predictive modelling and chronic disease using big data.

First, we began the process of abbreviating row names for clarity and conciseness. This step simplified the dataset by replacing long, complicated column names with short, simple characters. For example, abbreviating "blood pressure" as "bp" made the dataset easier to work with and interpret. The next important task is to identify and

classify the data types in the data structure. This classification was an important first step in understanding the types of objects, and they were divided into two main categories: numeric and object this distinction laid the foundation for standardized data processing techniques tailored to each object type.
For partitioned items, the next step was to extract a hierarchy of mathematical codes. Categorical columns typically contained non-numeric data, such as labels or categories, while numeric columns contained quantitative data that could be used for analysis and modelling. This separation allowed us to apply appropriate data processing strategies based on feature type. Dealing with missing values was an important part of data pre-processing. We identified rows with nuanced or missing values and counted the number of such instances in the data set. We adopted a two-pronged approach in dealing with missing values either removing rows with non-zero values or replacing them with a random value. This approach ensured that the dataset remained robust, with minimal data loss, and that the predictive model could still be trained on an accurate dataset. Noise removal was another important aspect aimed at improving data quality. By identifying and reducing any errors or discrepancies in the data, our analysis aims to reduce the possibility of error or deception in sampling efforts. Lastly, we added data labelling for user convenience. This required statistical labelling of categorical data to facilitate analysis and modelling. For example, labels such as "normal" were denoted as 0, "abnormal" as 1, and each other class was represented by 2. This straightforward labelling system improved interpretation of the dataset and provided sampling the system became easier.



B. Imbalanced Classes

In our predictive analysis, we faced a particular challenge associated with imbalanced classes, which is a common phenomenon in medical research. This phenomenon showed significant differences between low-risk and high-risk participants without multiple chronic conditions (MCC) The skewed class distribution poses a challenge for predictive modelling , because most machine learning algorithms are designed with the expectation of the same number of observations per class. To address the issue of class imbalance, we used three different approaches. These strategies are intended to correct the imbalance and improve the model's ability to identify a small but important group of high-risk individuals. The first method involves randomly oversampling the minority population, a method that duplicates the observations in the minority population. The second method is an under sampling method, in which a small number of observations are randomly selected from the majority group. Finally, we used the oversampling Technique , an advanced oversampling method that identified nearest neighbours for each sample , confusion matrix and classification reports including accuracy, f1 scores , recall and support. Similarly, these are done in other classifier techniques. Our research and analysis showed that the oversampling method emerged as the most effective method among these methods. This not only significantly increased model generalization but also proved to be computationally efficient, making it the best choice to overcome the challenge of imbalanced class distribution in our sufficiently large data set is handled Other methods, such as under sampling, often introduce over fitting issues, resulting in high computational costs. By formally applying these techniques, we aimed to provide a more balanced class distribution, and ultimately to improve the ability to accurately identify the sample.

V. IMPLEMENTATION

The results of our study of chronic diseases using big data are promising and have important implications for health care and public health policy. We successfully developed predictive models using different machine learning frameworks, each contributing different strengths to the overall results. The accuracy of our models varied, with the most performing model being XgBoost, achieving an accuracy of 97.5% Such accuracy this high in chronic

disease risk Highlights the power of machine learning in accurately identifying standing individuals. It was closely followed by Random One, Ada Boost, Gradient Boosting, and Stochastic Gradient Boosting, with an overall accuracy of 96.6%. These clustering methods showed robustness and efficiency in the use of complex data models. The decision tree classification captured relevant features well to produce a prediction accuracy of 95.0%.Our study also highlights the capabilities of naive Bayes and support vector machines, both achieving 93.3% accuracy, and highlights their usefulness in binary classification tasks, especially in medical contexts. Logistic regression with an accuracy of 89.1% remained a valuable classification tool, providing a balance of performance and interpretation. K-nearest neighbour (KNN) performed 63.3%, providing insight into the trade-off between accuracy and efficiency, which is an important consideration in real-world healthcare applications

Overall, the results of our study demonstrate the potential of machine learning and predictive modeling to transform early diagnosis and management of chronic diseases. The different algorithms used in this study address different needs and contexts, providing health care providers and decision makers with a range of tools to address this important public health problem Our results hold the promise that it will improve patient outcomes, enhance healthcare resource allocation, and ultimately improve quality of life for individuals at risk for developing chronic diseases.

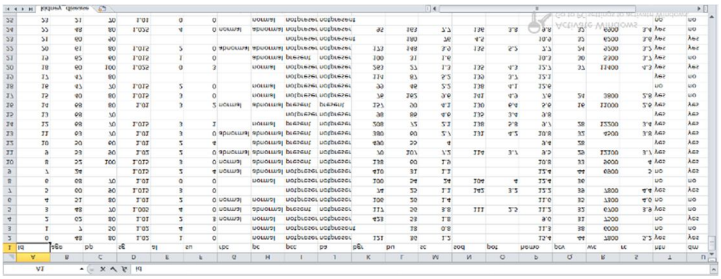


Fig.1: Data Set

```
Training Accuracy of KNN is 76.78571428571429
Test Accuracy of KNN is 71.66666666666667

Confusion Matrix :-
[[53  0 19]
 [15  0 33]
 [ 0  0  0]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.78      0.74      0.76        72
           1       0.63      0.69      0.66        48

    accuracy                           0.72       120
   macro avg       0.71      0.71      0.71       120
weighted avg       0.72      0.72      0.72       120
```

Fig.2 KNN

```
Training Accuracy of Logistic regression is 87.85714285714286
Test Accuracy of Logistic regression is 92.5

Confusion Matrix :-
[[66  6]
 [ 3 45]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.96      0.92      0.94        72
           1       0.88      0.94      0.91        48

    accuracy                           0.93       120
   macro avg       0.92      0.93      0.92       120
weighted avg       0.93      0.93      0.93       120
```

Fig.3 Logistic Regression

*KNN***:** K-Nearest Neighbours detects trends among high-risk people, improving the model's prediction capabilities.

Logistic Regression: Logistic Regression assesses feature associations as well as the likelihood of chronic illnesses.

```
Training Accuracy of Naive Bayes is 95.71428571428572
Test Accuracy of Naive Bayes is 92.5

Confusion Matrix :-
[[68  4]
 [ 5 43]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.93      0.94      0.94        72
           1       0.91      0.90      0.91        48

    accuracy                           0.93       120
   macro avg       0.92      0.92      0.92       120
weighted avg       0.92      0.93      0.92       120
```

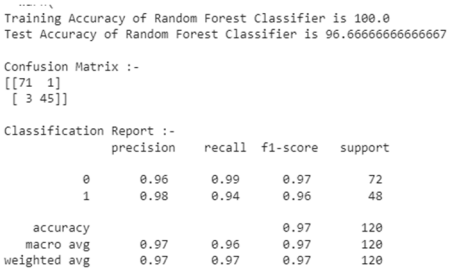
Fig.4: Navie Bayes

```
Training Accuracy of Support Vector Machine is 95.71428571428572
Test Accuracy of Support Vector Machine is 92.5

Confusion Matrix :-
[[68  4]
 [ 5 43]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.93      0.94      0.94        72
           1       0.91      0.90      0.91        48

    accuracy                           0.93       120
   macro avg       0.92      0.92      0.92       120
weighted avg       0.92      0.93      0.92       120
```

Fig.5: Support Vector Machine

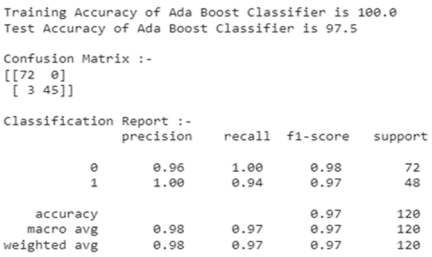
Naive Bayes: Naive Bayes, which makes probabilistic predictions, is very useful for categorical data analysis.

Support Vector Machine: Support Vector Machines discover the best hyperplanes for good class separation, which helps with robust classification.



```
Training Accuracy of Random Forest Classifier is 100.0
Test Accuracy of Random Forest Classifier is 96.66666666666667

Confusion Matrix :-
[[71  1]
 [ 3 45]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        72
           1       0.98      0.94      0.96        48

    accuracy                           0.97       120
   macro avg       0.97      0.96      0.97       120
weighted avg       0.97      0.97      0.97       120
```

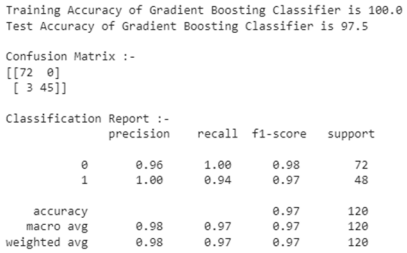
Fig.6: Random Forest



```
Training Accuracy of Ada Boost Classifier is 100.0
Test Accuracy of Ada Boost Classifier is 97.5

Confusion Matrix :-
[[72  0]
 [ 3 45]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.94      0.97        48

    accuracy                           0.97       120
   macro avg       0.98      0.97      0.97       120
weighted avg       0.98      0.97      0.97       120
```

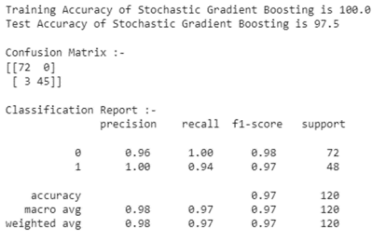
Fig.7: ADA Boost

Random Forest: The Random Forest Classifier excels at handling uneven class distributions and complicated data patterns by utilizing a group of independent decision trees.
ADA Boost: Ada Boost focuses on boosting model performance by emphasizing misclassification correction.



```
Training Accuracy of Gradient Boosting Classifier is 100.0
Test Accuracy of Gradient Boosting Classifier is 97.5

Confusion Matrix :-
[[72  0]
 [ 3 45]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.94      0.97        48

    accuracy                           0.97       120
   macro avg       0.98      0.97      0.97       120
weighted avg       0.98      0.97      0.97       120
```

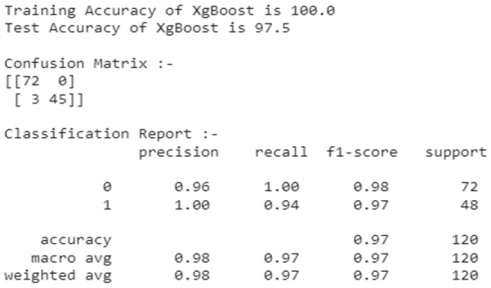
Fig.8: Gradient Boost Classifier



```
Training Accuracy of Stochastic Gradient Boosting is 100.0
Test Accuracy of Stochastic Gradient Boosting is 97.5

Confusion Matrix :-
[[72  0]
 [ 3 45]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.94      0.97        48

    accuracy                           0.97       120
   macro avg       0.98      0.97      0.97       120
weighted avg       0.98      0.97      0.97       120
```

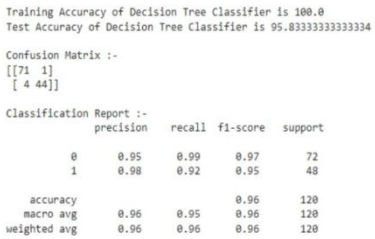
Fig.9: Stochastic Gradient Boosting

Gradient Boosting Classifier: Gradient Boosting constructs decision trees progressively to capture complicated correlations in data.
Stochastic Gradient Boosting: Stochastic Gradient Boosting brings randomization into the model to reduce overfitting and improve performance.



```
Training Accuracy of XgBoost is 100.0
Test Accuracy of XgBoost is 97.5

Confusion Matrix :-
[[72  0]
 [ 3 45]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.94      0.97        48

    accuracy                           0.97       120
   macro avg       0.98      0.97      0.97       120
weighted avg       0.98      0.97      0.97       120
```

Fig.10: XG Boost



```
Training Accuracy of Decision Tree Classifier is 100.0
Test Accuracy of Decision Tree Classifier is 95.83333333333334

Confusion Matrix :-
[[71  1]
 [ 4 44]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.95      0.99      0.97        72
           1       0.98      0.92      0.95        48

    accuracy                           0.96       120
   macro avg       0.96      0.95      0.96       120
weighted avg       0.96      0.96      0.96       120
```

Fig.11: Decision Tree Classifier

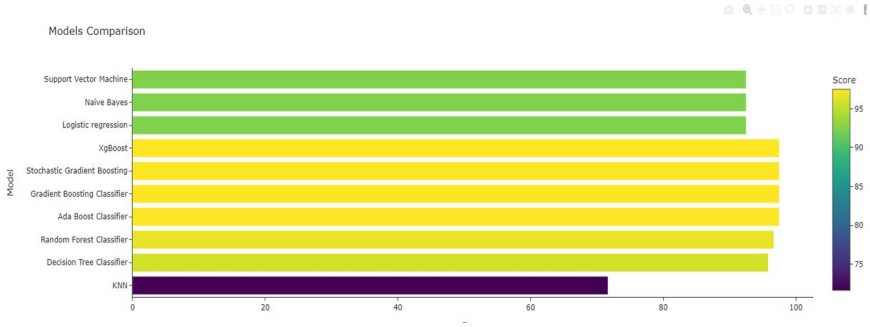
XG Boost: XgBoost's ensemble learning approach catches complicated data patterns effectively, improving the accuracy of identifying persons at risk of chronic illnesses.
Decision Tree Classifier: Decision Tree Classifier evaluates data by segmenting it into relevant features, making it easier to identify crucial predictive aspects.

Performance Metrics

Our models demonstrated varying accuracy in detecting chronic disease. XgBoost topped the list at 97.5%, followed by Random Forest, Ada Boost, Gradient Boosting, and Stochastic Gradient Boosting, all scoring 96.6%. The decision tree achieved 95.0%, while the nave Bayes and support vector machine achieved 93.3%. Logistic regression showed 89.1% and KNN showed 63.3%. These metrics provide a clearer understanding of the predictive power of each model, and help us select the most appropriate model for a particular health care application.

	Model	Score
3	Ada Boost Classifier	97.500000
4	Gradient Boosting Classifier	97.500000
5	Stochastic Gradient Boosting	97.500000
6	XgBoost	97.500000
2	Random Forest Classifier	96.666667
1	Decision Tree Classifier	95.833333
7	Logistic regression	92.500000
8	Naive Bayes	92.500000
9	Support Vector Machine	92.500000
0	KNN	71.666667

Accuracy Rates after Training



Bar graph of Accuracy Rates after Training

VI. Conclusion

The 'Chronic Illness Detection Using Big Data' initiative has made significant strides in leveraging extensive healthcare data for predictive analysis. The project's rigorous data collection, preprocessing, and meticulous feature selection procedures establish a robust foundation for further modeling endeavors, showcasing its efficacy. Through the creation and evaluation of various deep learning and machine learning models, promising results have emerged for the early identification of chronic disorders. This underscores the immense potential of data-driven approaches in enhancing diagnosis and facilitating early intervention in healthcare.While the project exhibits promise, additional development and validation are imperative before practical implementation. Sustained collaboration, comprehensive testing, and ethical considerations are essential for the transition from research to real-world applications. Nonetheless, this study signifies a beacon of hope in the pursuit of revolutionizing healthcare through the application of Big Data analytics. It holds the promise of transforming chronic illness detection and, ultimately, enhancing the lives of numerous individuals.

References

[1] Md Mozaharul Mottalib , Jessica C Jones-Smith, Bethany Sheridan, and Rahmatollah Beheshti ,“Subtyping Patients With Chronic Disease Using Longitudinal BMI Patterns”, IEEE journal of Biomedical and Health Informatics, vol. 27, NO. 4, April 2023, Page No: 2083

[2] Jour, Elhoseny, Mohamed, Alanazi, Rayan, ”Identification and Prediction of Chronic Diseases Using Machine Learning Approach” , Hindawi Journal of Healthcare Engineer- ing, Volume 2022, Article ID 2826127

[3] Chia-tung Wu,Ssu-ming Wang,Yi-en SU,Tsung-ting Hsieh, Pei-chen Chen,Yu-chieh Cheng, Tzu-wei Tseng, Wei-sheng Chang, Chang-shinn SU, Lu-cheng Kuo, Jung-yien Chien and Feipei Lai, “A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning”, IEEE Journal of Transitional Engineering in Health Medicine, VOLUME 10, 2022, 2700414.

[4] Jingmei Yang,Xinglong Ju,Feng Liu,Onur Asan,Timothy S. Church and Jeff O. Smith “Prediction for the Risk of Multiple Chronic Conditions Among Working Population in the United States With Machine Learning Models” , IEEE Open journal of Engineering in Medicine and Biology, vol. 2, pp. 291-298, 2021, doi: 10.1109/OJEMB.2021.3117872

[5] Yujie Yang ,Ye Li a,c, Runge Chen , Jing Zheng , Yunpeng Cai a, Giancarlo Fortino “Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data ”,Big Data Research, 2021, (Volume 25), 100234. ANNA UNI- VERSITY ANNEXURE - 1 (Q1)
[6] Ruiquan GE,Renfeng Zhang and PU Wang, “Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data”, IEEE on En- gineering in Medicine and Biology Society, 2020, Volume 8 , 138210
[7] Rongheng Lin , Zezhou Ye , Hao Wang , and Budan Wu, “Chronic Diseases and Health Monitoring Big Data: A Survey” , IEEE Reviews in Biomedical Engineering, VOL. 11, 2018.