CENTER OF EXCELLENCE FOR OPEN SCIENCE

Open Access

Email: cphr@ceospublishers.com

CEOS Public Health and Research

Research Article

Machine Learning for Chronic Disease Prediction

Al Khan*

Ala Too International University, Kyrgyzstan

Received Date: July 02, 2022
Accepted Date: August 02, 2022
Published Date: August 05, 2022

***Corresponding Author**

Al Khan, Ala Too International University, Kyrgyzstan. Tel: 996999077221, Email: pkhan_1@hotmail.com

Citation

Al Khan (2022) Machine Learning for Chronic Disease Prediction. CEOS Public. Health. Res. 1(1):101

Abstract

As a result of current environmental circumstances and lifestyle choices, humans are increasingly prone to a number of diseases. Early detection and prognosis of such disorders is crucial for avoiding their worst manifestations. The majority of the time, doctors find it difficult to diagnose illnesses accurately by hand. The goal of this study is to identify and predict persons who are more likely to develop chronic illnesses. This might be achieved by utilizing a cutting-edge machine learning technique to reliably identify patients with chronic illnesses. Predicting disease is also a challenging task. As a result, data mining is critical for predicting sickness. The proposed system uses machine learning algorithms such as convolutional neural network (CNN) for automatic feature extraction and disease prediction and K-nearest neighbor (KNN) for distance calculation to find the exact match in the data set and the final disease prediction outcome to provide a broad disease prognosis based on the patient's symptoms. For the development of the data set, a collection of sickness symptoms, as well as the person's living habits and details linked to doctor consultations, were collected. Finally, this work presents a comparison of the proposed system with several methods such as Nave Bayes, decision trees, and logistic regression.

Keywords: Convolutional Neural Network (CNN), K-Nearest Neighbor (KNN), Precision, Accuracy, Nave Bayes, Decision Trees, Logistic Regression

Introduction

Chronic illnesses are a major problem in the healthcare industry all around the world. According to the medical statement, the death rate of people is increasing as a result of chronic disorders. The cost of treating this condition consumes more than 70% of the patient's income. As a result, it is critical to reduce the patient's risk of mortality. The development of medical research has made it simpler to acquire health-related data [1]. The demographics, medical analysis reports, and the patient's disease history are all included in the healthcare data. The illnesses that result might differ depending on the place and the living environments found there. As a result, in addition to illness data, the patient's environmental state and living environment should be documented in the data set. The integration of information technology (IT) into the healthcare area has accelerated its evolution in recent years. The goal of integrating IT into healthcare is to make people's lives more inexpensive and comfortable, similar to how smartphones have made people's lives simpler [2]. This might be accomplished by making healthcare smarter, such as through the development of smart ambulances, smart hospital facilities, and other innovations that benefit patients and clinicians in a variety of ways [3]. Every year, a study of patients with chronic diseases in a specific location was conducted, and it was discovered that the gender gap between patients is extremely tiny, and that a high number of patients were hospitalized for treatment of chronic diseases in 2014. Instead of using solely structured data, using both structured and unstructured data yields very accurate outcomes. Because unstructured data includes doctor's records on patients with diseases, as well as the patients' symptoms and grievances, which can be explained by the patients themselves, it has a distinct advantage when combined with structured data, which includes patient demographics, disease details, living habitats, and laboratory test results [3]. Rare illnesses are difficult to diagnose. As a result, using self-reported behavioral data can assist distinguish those with unusual diseases from those with typical chronic conditions. The detection of uncommon illnesses is thought to be highly doable utilizing machine learning algorithms combined with questionnaires [4].

MRI (magnetic resonant imaging) readouts, ultrasonography, social media data, and electronically obtained activity, behavioral, and clinical data were all launched in the recent decade as revolutionary tools to quickly collect data. These healthcare large data sets are multidimensional, which implies the number of characteristics recorded each observation may exceed the total number of observations. They're noisy, scarce, cross-sectional, and statistically insignificant. Machine learning approaches can be used to solve problems with high-dimensional data sets [5]. Machine learning is more useful in a variety of fields. Many of the complicated models rely on existing bigger training data, which comes at a time when healthcare epidemiology is undergoing a dramatic transition. These data can help researchers learn more about disease risk factors, which can help them minimize healthcare-associated infections, improve patient risk stratification, and figure out how infectious diseases spread [6]. Machine learning can help with the examination of test data and other patient information for illness early detection. Low-level data might be turned to high-level information using database knowledge discovery to learn about illness trends and promote early detection. To improve prediction accuracy while reducing model training time, the data gathered for constructing a data set should be preprocessed for missing values, and then just the key features needed for successful illness prediction should be picked.

People are unconcerned about their health and life in the age of the Internet and technology. Because everyone is preoccupied with surfing and social media, they neglect to attend clinics for a health exam. Using this activity as a benefit, a machine learning model should be constructed that takes the symptoms as input and forecasts the likelihood and risk of the illness being impacted or the development of such diseases in an individual. Diabetes, cardiovascular disease, cancer, strokes, hepatitis C, and arthritis are the most frequent chronic illnesses. The identification of these disorders is critical in the healthcare sphere since they last a long time and have a high death rate. Predicting the condition can help you take preventative measures and avoid becoming ill, and early discovery can help you obtain better treatment. Machine learning includes supervised, semi-supervised, unsupervised, reinforcement, evolutionary, and deep learning approaches. The issue is linked to the processing of extracted characteristics from real-world data that are organized as vectors [7]. The right mix of such vectors determines the processing quality. However, the large dimensionality of the vectors or disparities in the data are frequently a problem. As a result, it's critical to lower the data set's dimensionality, even if it means losing certain information, in order to make it a highly compatible dimension.

The model's performance is improved by reducing the data set's dimensionality [8].

For people who are affected by chronic illnesses and require adequate medical evaluation and treatment information, the chronic disease management system is critical. This technique can also be beneficial for people who need self-care to enhance their health, since it has been proven that self-management is the primary care for people with chronic conditions, and it is an inescapable element of treatment. Patients' health information may be captured via mobile applications, and they serve as a superior tool for enabling self-management. Information such as patient narration of symptoms, details of consultations with medical practitioners, lab examination findings, and computed tomography and X-ray pictures are used to efficiently forecast illness [9]. There has been little study into determining the accuracy and predictive capacity of constructing a machine learning model for illness detection using simply information from lab test findings. Ensemble machine learning and deep learning models can also be employed to improve performance [10]. In the healthcare domain, artificial intelligence (AI) plays a major role in automating the roles involved in disease diagnosis and treatment suggestions and also schedules perfect timing by the medical practitioners to perform various obligations that cannot be automated.

The suggested system's main goal is to use a machine learning technique to identify and forecast chronic illness in an individual [11]. The data set includes both structured and unstructured data, which includes the patient's age, gender, height, weight, and other demographic information (excluding personal information such as name and ID) and unstructured data, which includes the patient's symptoms, information related to disease consultations with doctors, and the patient's living habits. The missing values are found using preprocessed data. They are then rebuilt to improve the model's quality and, as a result, the forecast accuracy. Machine learning techniques such as CNN and KNN are utilized for prediction [12]. This paper is organized as follows: Section 2 contains details of related works completed while conducting the research, Section 3 contains preliminaries of the algorithms used, Section 4 contains a description of the proposed methodology, Section 5 contains the results and discussion section, Section 6 contains the conclusion, and Section 7 contains a list of references used in this study.

Related Work

This section summarizes the associated research that went into constructing the suggested model for chronic illness prediction. The following are the conclusions reached after analyzing the current literature, which aid in the efficient and successful development of the suggested system.

The study's target variable is resource consumption, such as medical and long-term care costs, as well as a medical care forecasting model utilizing a random forest machine learning algorithm. This strategy employs over 100 pieces of information, including preventative behaviors, clinical testing, and medicinal procedures. For classification, the mean reduction Gini is employed, and for regression, the mean square error (MSE) is utilized [13]. The training model employs grid search for hyper parameter tweaking and K-fold cross-validation for validation. Because the goal of this article is to properly manage the budget for medical treatment, exploratory factors such as age, gender, and analysis period are included alongside the objective variable. [14] provides a review of the applications of machine learning techniques in various medical practices, including predicting, diagnosing, and prognosis of diseases like multiple sclerosis, autoimmune chronic kidney disease, autoimmune rheumatic disease, and inflammatory bowel disease, as well as treatment selection and stratification of patients; drug development; drug repurposing; target interpretation; and validation. This study also goes through the obstacles that machine learning techniques confront, such as the necessity for high-quality data in the construction of robust models, external model validation using a separate data set, difficulties encountered during model implementation, and ethical considerations. [15] explains a prediction model for chronic renal disease. For classification purposes, this model is built utilizing four machine learning approaches: support vector machine (SVM), logistic regression (LR), decision tree (DT), and KNN. The data set utilized in this research is the Indian chronic kidney disease (CKD) data set, which has 400 occurrences, 24 characteristics, and two classes and can be found in the UCI machine learning repository. The created model is tested using a 5-fold cross-validation procedure, and the experiment is carried out using the Weka data mining tool and MATLAB, with the conclusion that the SVM classifier achieves greater accuracy than the others.

[16] describes a system that can predict different illnesses using various machine learning techniques such as Nave Bayes, KNN, DT, random forest, and SVM algorithms to bridge the gap between patients and physicians in order to reach their own goals. Existing efforts in the field of automatic illness prediction lack patient trust in the model's predictions and minimize the need for physicians, causing doctors to become concerned about their livelihood. However, this system incorporates a module for doctor suggestion, which addresses both difficulties by ensuring patient confidence through doctor intervention while also improving physicians' revenue. In [17], a model named PARAMO was built, which is a framework for a parallel predictive model that employs electronic health records (EHR) for healthcare analysis. The development of the dependency graph, which removes duplication and identifies dependencies, is followed by the execution engine for the dependency graph, which includes prioritization, scheduling, and parallel execution, and lastly the parallelization infrastructure. However, this system incorporates a module for doctor suggestion, which addresses both difficulties by ensuring patient confidence through doctor intervention while also improving physicians' revenue. The PARAMO model is tested with three sets of actual data: small, medium, and big data sets, which comprise prescriptions, diagnostic data, and lab reports and are gathered from EHRs with patient populations ranging from 5,000 to roughly 300,000. In addition, the small and big sets contain procedure data, whereas the medium set contains heart failure symptoms extracted from medical records [18]. [19] established an effective suggestion method for chronic illness diagnosis. A data mining strategy is used in this procedure. Medical data and two-dimensional data are among the data sets employed in this system. The two-dimensional data includes the exterior user and item attributes, and the medical data includes data acquired from sensors or medical data entries. The decision tree technique, which is a widely used data mining approach, is used for categorization to improve prediction accuracy. This prediction model uses a variety of decision tree classifiers, including random forest, REP tree, decision stump, and J48. A data mining strategy is used in this procedure. Medical data and two-dimensional data are among the data sets employed in this system. The two-dimensional data includes the exterior user and item attributes, and the medical data includes data acquired from sensors or medical data entries. The decision tree technique, which is a widely used data mining approach, is used for categorization to improve prediction accuracy. This prediction model uses a variety of decision tree classifiers, including random forest, REP tree, decision stump, and J48. The RF method outperforms the other three algorithms when evaluated with 20 randomly selected samples.

[20] shows how decision trees, maximum margin learning, and instance-based learning may be used to predict three types of immunological disorders: allergy, infectious, and autoimmune diseases. One of the goals of this research is to find a link between immunogen categorization and their physicochemical features. The Immune Epitope Database (IEDB) was used to collect immunogen data such as disease statistics, B-cell responses, discontinuous epitope location, host, source organisms, and so on, and to analyze its six physicochemical properties, including PSSM (position-specific scoring matrix) information per position, hydrophilic scale, flexibility, antigenic propensity, hydropathy index, and side chain polarity. For the performance of prediction outcomes using metrics such as accuracy and F-score, this system is examined using a method known as leave-one-out cross-validation. [20] shows how decision trees, maximum margin learning, and instance-based learning may be used to predict three different types of immunological disorders, including allergy, infectious, and autoimmune diseases. One of the goals of this research is to examine the relationship between immunogen categorization and physicochemical parameters. Immunogen data, such as disease statistics, B-cell responses, discontinuous epitope location, host, source organisms, and so on, were gathered from the Immune Epitope Database (IEDB) and analyzed using six physicochemical properties, including PSSM (position-specific scoring matrix) information per position, hydrophilic scale, flexibility, antigenic propensity, hydropathy index, and side chain polarity. For the performance of prediction outcomes using criteria like accuracy and F-score, this system is assessed using a method known as leave-one-out cross validation.

[21] describes a risk prediction model for forecasting illness risks from highly imbalanced data that employs a random forest machine learning technique. The Nationwide Inpatient Sample (NIS), which contains 8 million records of hospital stays with 126 clinical and nonclinical variables, was used in this study. Patient demographics, hospital location, date and year of admission, pin code, treatment/diagnosis cost, and length of stay in a hospital ward are all included in the nonclinical data. The therapy processes, their categories, diagnostic categories, and codes make up the clinical data. Each record includes a vector

with 15 diagnostic codes from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Because the unbalanced data gives unwanted outcomes, the problem is solved via a repeated random sampling procedure. SVM, RF ensemble learning, bagging, and boosting techniques are used to assess the created model. The work [22] shows how to detect chronic kidney disease in its early stages using a unique adaptive probabilistic divergence-based feature selection technique. The statistical and divergent information theories are used to create this method.

In this work, the hyper parameterized logistic regression model is applied for classification. The information for 630 patients with 52 characteristics is acquired from multiple hospitals and laboratories, and the data set is delivered to the physician for verification. The precision, recall, F1-score, and ROC (receiver operating characteristics) curve are used to evaluate the model developed using data sets from diabetes, heart disease, and kidney disease, and the performance evaluation metrics used in this study are precision, recall, F1-score, and ROC (receiver operating characteristics) curve.

[23] shows a system that uses a deep learning technique on huge data and an updated fusion node model to improve risk prediction of a patient's health state. This deep learning model uses a combination of complicated machine learning algorithms like Bayesian fusion and neural networks to extract data and make logical inferences. This system's architecture is made up of five layers: a data layer for data collection, a data aggregation layer for data acquisition from multiple data sources and format conversion, an analytics layer for proper data analytics, an information exploration layer for creating output that makes the results of analytics understandable for users, and a big data governance layer for managing the above layers. Also explored in this article is the use of MapReduce to improve analytics performance and to inspire the design of SOA (service-oriented architecture) to allow other systems to easily access analytics findings. A machine learning disease prediction model created [24] is of the cost that leverages big data to prepare the data set, which comprises structured and unstructured data, and the produced model is made accessible at a low cost. The decision tree algorithm is employed as the prediction algorithm in this technique, and the MapReduce algorithm is used to improve the efficiency of the operation. The advantages of this paradigm include faster query retrieval and greater accuracy. [25] offer a strategy for forecasting the risk of chronic renal disease using zub machine learning algorithms. This strategy employs two different types of data sets. One comes from UCI and has 400 instances and 35 features, while the other comes from Khulna City Medical College and has 55 instances and 25 features. The Pandas and Numpy libraries are used to process the data, and median filtering is used to deal with missing data. The Chi-square test is used for feature extraction. The model is evaluated using 10-fold cross-validation. Disease classification techniques include artificial neural networks (ANN) and random forest algorithms. This approach [26] is thought to be able to forecast the likelihood of chronic kidney disease at an early stage.

Preliminaries

Chronic Disease

Chronic diseases, according to the National Center for Health Statistics in the United States, are illnesses that continue longer than three months. Neither medications nor vaccinations can cure or prevent these disorders. Tobacco use, bad eating habits, and a lack of physical activity are the leading causes of chronic illnesses. Furthermore, this condition is frequently induced by aging. Cardiovascular disease, cancer, arthritis, diabetes, obesity, epilepsy and seizures, and oral health issues are examples of chronic illnesses [27].

Heart disease and stroke are two types of cardiovascular illness that frequently result in mortality. Tobacco use, nutrient-deficient diets, and a lack of physical activity all contribute to this condition. When these behaviors are modified by the patient, the influence on managing and preventing cardiovascular disease may be reduced.

Cancer, such as colon cancer and breast cancer, is the second-deadliest illness after cardiovascular disease. Only prevention, early discovery, and competent medical care can keep it under control. Reduce the risk of cancer through reducing the prevalence of cancer-causing environmental and behavioral variables.

Arthritis is a chronic condition that causes inflammation in the joints, discomfort, and stiffness, which worsens with age. There are cost-effective techniques for minimizing the consequences of arthritis, but they are not widely employed. Regular moderate exercise can help to lessen the affects of arthritis.

Diabetes is a costly and life-threatening condition. Self-care and early identification of diabetes can lessen the disease's burden. This illness, particularly type 2 diabetes, affects around 7 million persons over the age of 65.

Obesity has been more prevalent in persons of all ages since 1980. Overweight or obese people are more likely to develop high blood pressure (BP), heart disease, diabetes, and arthritis. Obesity has also been linked to the development of some malignancies.

Treatment for epilepsy and seizures is quite expensive. This condition affects people of all ages, but notably children and the elderly.

Oral health issues are a critical issue that receives extra attention in the health of the elderly. This is a major problem since it interferes with a person's ability to talk, chew, swallow, and follow a healthy diet plan.

Convolutional Neural Network (CNN)

The ConvNet or CNN, as indicated in Algorithm 1, is a deep learning algorithm that receives input, assigns bias and weights to its numerous features, and then differentiates one from the other [28]. The key benefit of using CNN over other algorithms is that it requires less effort in data preparation since it can learn to enhance filters through automatic learning [29]. The CNN output layer may be calculated using the equation as given:

Equation 1

Output height = (Input height + padding height top + padding height bottom - kernel height) / (stride height) + 1. Output width = (Output width + padding width right + padding width left - kernel width) / (stride width) + 1

K-Nearest Neighbor (KNN)

As illustrated in Algorithm 2, KNN is a supervised machine learning algorithm that evaluates the similarities between new and existing data and adds the new data to a category that is very comparable to the current categories [30]. The KNN may be used in both classification and regression tasks, however classification is the most typical use. Because it will not learn quickly from the training data, this algorithm is also known as the lazy learner algorithm. It saves the data set and performs its action throughout the classification process. The formula for calculating Euclidean distance is as follows:

distance between two points: (x1 1, y1 1) and (x2 2, y2 2) is $\mathbf{d} = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

Proposed Methodology

A full description of the data set development, model setup, and the disease prediction has been provided in this part. The first step is to gather information. Our suggested system gathers both structured and unstructured data from a variety of sources. Following data collection, they are preprocessed and divided into cleaning and test data sets. The training data set is then trained over a number of trials using machine learning techniques such as CNN and KNN to improve the accuracy of the prediction findings. Once the intended objective has been reached after numerous attempts, the produced model is ready for testing.

Data Collection: The actual required data contains both structured and unstructured data, such as the patient's basic demographics, living environment, and lab test results, as well as unstructured data, such as the patient's symptoms of the condition and their consultation with the doctor. To protect the patient's privacy, the data set eliminates personal information such as name, ID, and location.

Pre-Processing: In most structured data, the gathered data is preprocessed for the presence of missing values. As a Result, missing data must be filled in, or data must be removed or modified to improve the data set's quality. The commas, punctuations, and white spaces are also removed during the preprocessing stage. After the data has been preprocessed, it is subjected to feature extraction, followed by disease prediction.

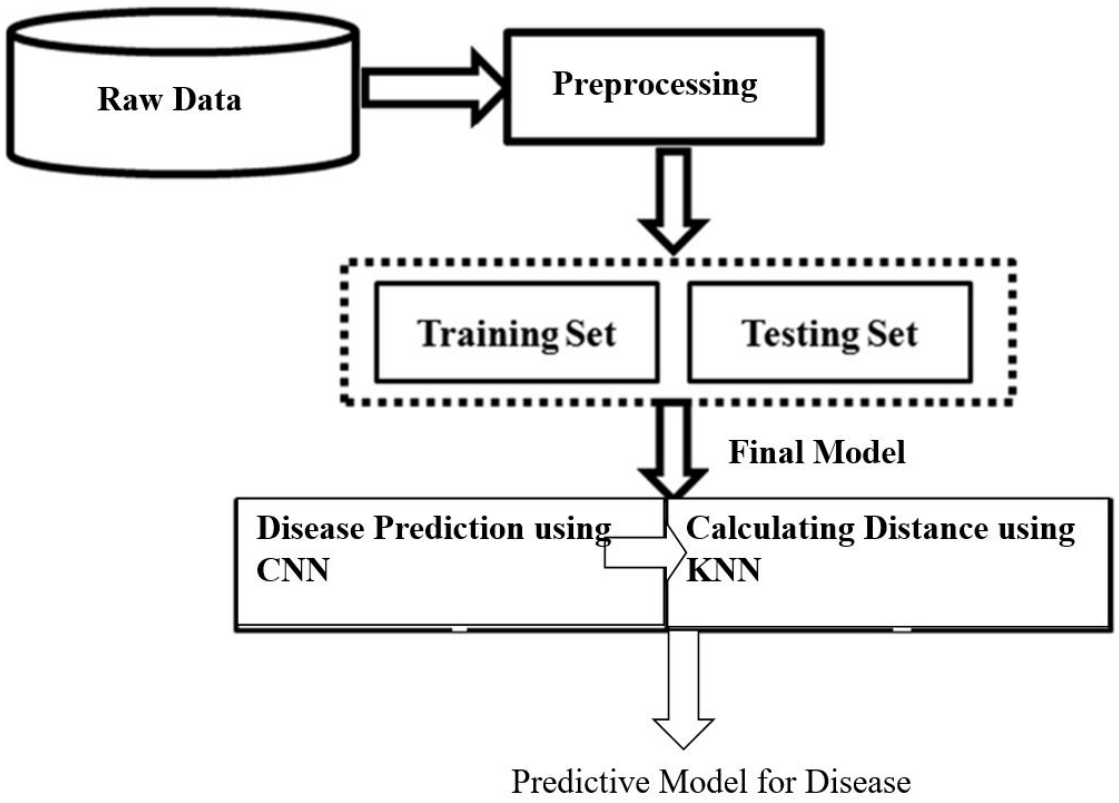


Figure 1: Proposed model of the disease predicting system

Model Description

The data set contains both structured and unstructured data, as stated above. In tabular format, the structured data includes patient demographics and data relevant to the disease's etiology, such as age, gender, height, weight, and so on, as well as the patient's living environment, laboratory test results, and the disease to which they are exposed. The unstructured data includes the patient's medical symptoms as well as information on the doctor's inquiry in text format. The use of unstructured data in the prediction task improves the accuracy of the outcomes. The data set is divided into two parts: 80 percent for training and 20% for testing.

Disease Prediction Using CNN: In order to forecast chronic diseases, the proposed methodology employs the CNN algorithm. The data set is first transformed to vector form, then word embedding is used to fill the data with zero values. And then it is passed to the convolution layer.

The pooling layer receives the convolution layer's input and applies the maximum pooling procedure. The fully connected layer receives the output of max pooling, and the output layer offers the classification results.

Distance Calculation Using KNN: The value of K is known in K-Nearest Neighbor (KNN), and the characteristics that are comparable to the K value are called the nearest neighbors. The nearest neighbor to the known K value is picked, and their distance is determined. The precise match, which is the final illness prediction result, is the characteristic with the lowest distance value. The suggested system employs Euclidean distance since its results are superior than those produced by other distance computation methods. It's a nonparametric algorithm since it doesn't make judgments based on raw data. The training input data in KNN are plots of X and Y axes, while the test data are in the plots of X and Y axes. The plots of test data with the shortest distance are then chosen as the intended goal. It is critical that the value of the closest K point be odd at all times.

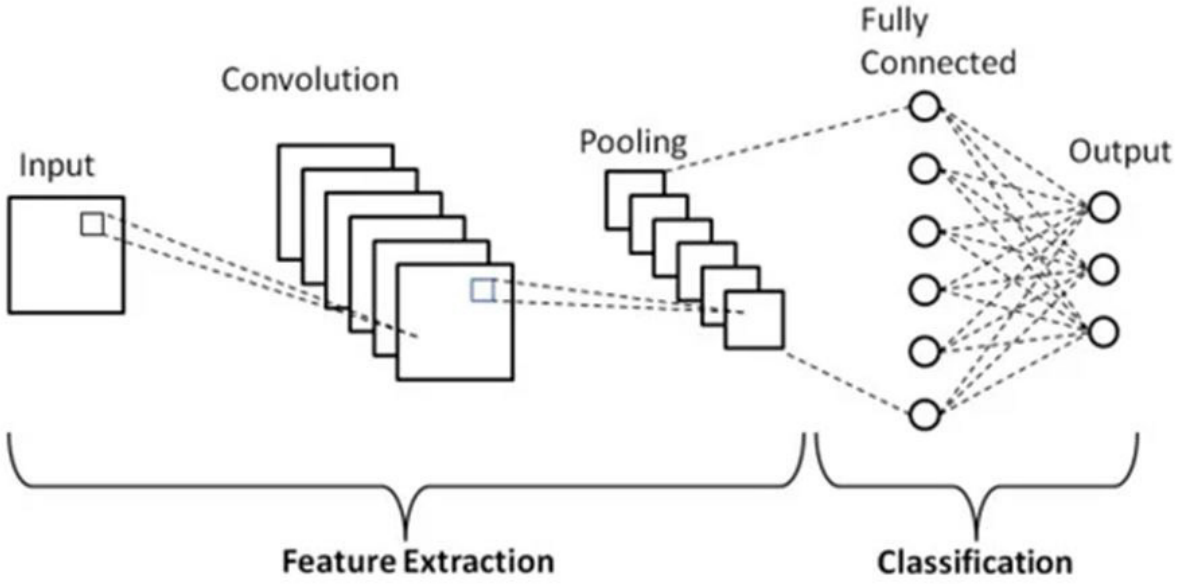


Figure 2: The model of the convolutional neural network

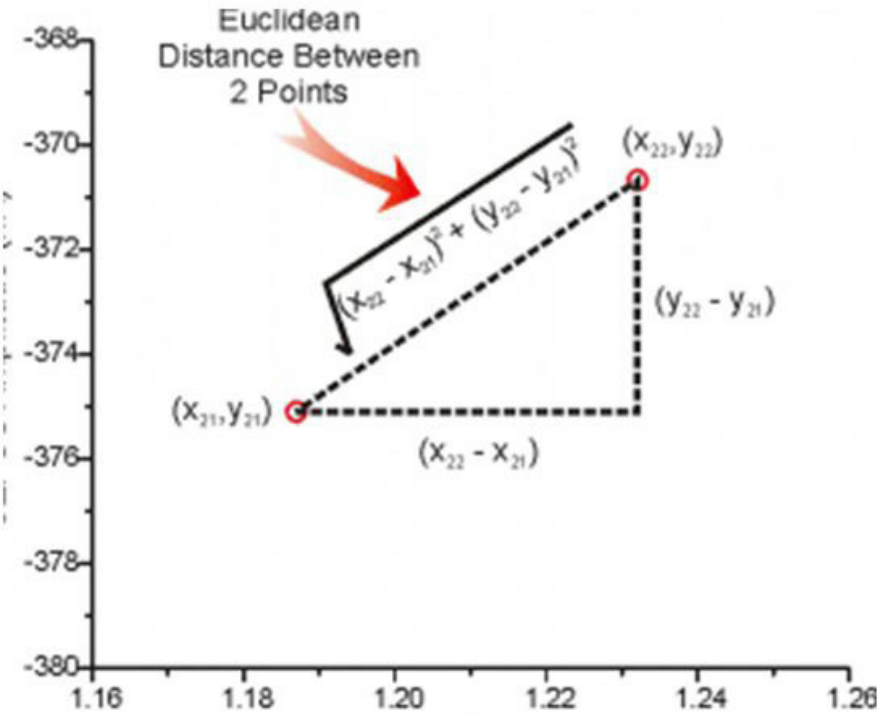
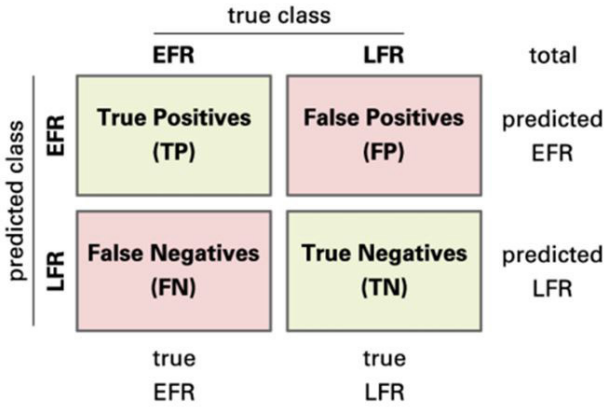


Figure 3: Represents the formula to Calculation of Euclidean distance

3) Represents the formula to calculate the Euclidean distance

Equation 1



$$PR = \frac{TP}{TP+FP}$$

$$RE = \frac{TP}{TP+FN}$$

$$CA = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F_1 = \frac{2TP}{2TP+FP+FN}$$

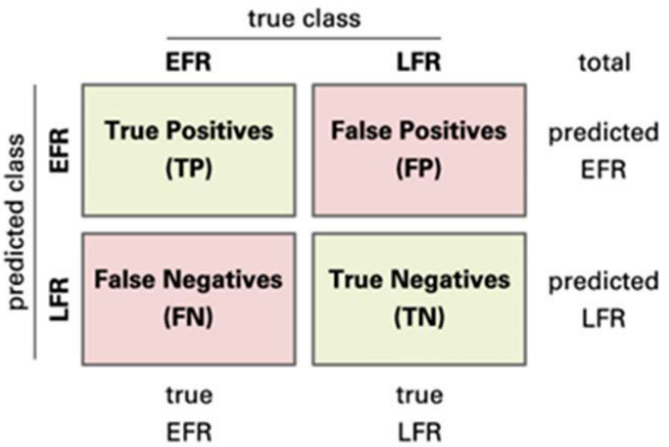
Model Evaluation

Following the training and construction of a prediction model, its accuracy is evaluated by testing it on comparable characteristics with an unknown output class. To verify the resemblance between the actual output and the proposed model's predictions, a confusion matrix is created. True positive (TP) indicates that a positive output is predicted as positive, false positive (FP) indicates that a negative class is predicted as a positive class, true negative (TN) indicates that a negative class is correctly predicted as a positive class, and false negative (FN) indicates that a negative class is incorrectly predicted as a positive class in the confusion matrix. Any categorization model's key measures are these. Accuracy is one of the most commonly utilized performance indicators. It's the proportion of correct forecasts to total cases. The four performance assessment parameters are described below.

Accuracy:

The classification accuracy is expressed mathematically as the proportion of accurate predicted values to total predicted values: The accuracy formula is represented in Equation 2:

Equation 2



D = √[(X2-X1)^2 + (Y2-Y1)^2)

- Where D is the distance
- X1 and X2 are the x-coordinates
- Y1 and Y2 are the y-coordinates

$$F_1 = \frac{2TP}{2TP+FP+FN}$$

Precision:

The precision, also known as the positive predictive value (PPV), is defined as the ratio of correct forecasts to total correct values, which includes both true and erroneous predictions and is represented above in the equation 2.

Recall:

The recall, sensitivity, or true positive rate (TPR) is defined as the ratio of correct predicted values to the sum of correct positive and incorrect negative predicted values, as shown in the above equation 2.

F1-Score:

The weighted average of the values obtained from the calculation of accuracy and recall parameters is known as the F-measure (Fϐ).

When the distribution of class is not even, the F1 Score value is more essential than the accuracy number. When the values of false positives and negatives are different, the F1 Score value is an excellent choice. The above in equation 2 is a mathematical representation of the F1 Score.

The suggested CNN and KNN model's accuracy, recall, and F1-score are compared to the performance metrics of the Nave Bayes, decision tree, and logistic regression methods, and the results are reported in Table 1. Because the forecast result is such a crucial component for the patient, and if it is incorrect, it would be detrimental to them, accuracy is a critical parameter.

The other parameters in Table 1 are precision, recall, and F1-score, which are used to evaluate the model's performance.

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Na¨ıve Bayes	52	52	80	65
Decision tree	62	64	60	62
Logistic regression	86	84	88	82
CNN and KNN	96	93	99	97

Table 1: Performance evaluation comparison

Classifier	Train_Score	Test_Score	Training_time
Naïve Bayes	0.7672	0.7619	0.0041
Logistic Regression	0.7672	0.7836	0.0190
Random Forest	0.9963	0.7706	0.1146
K Nearest Neighbors	0.7896	0.7489	0.0030
Decision tree	1.0000	0.7403	0.0079

Table 2: Comparison of Algorithms for training time and Score

Figure 4 depicts the graphical depiction of the proposed and alternative algorithms' comparison precision, recall, and F1-score values. This graph depicts the variations in the three performance evaluation parameters of the four algorithms: 52 percent, 64 percent, 84 percent, and 93 percent, respectively, for precision; 80 percent, 605, 88 percent, and 99 percent, respectively, for recall; and 65 percent, 62 percent, 82 percent, and 97 percent, respectively, for F1-score. These findings reveal that the prosed model constructed using the CNN and KNN algorithms is the best of the remaining three algorithms, with accuracy, recall, and F1-score of 93 percent, 99 percent, and 97 percent, respectively, which is greater than the others.

Figure 5 as given above, gives a graphical depiction of the accuracies of the proposed and other methods in comparison. The prediction accuracies of the four algorithms (Nave Bayes, decision tree, logistic regression, and the suggested CNN and KNN algorithms) vary by 52 percent, 62 percent, 86 percent, and 96 percent, respectively, as shown in this graph. When compared to existing machine learning methods, the suggested approach obtains the maximum accuracy of 96 percent.

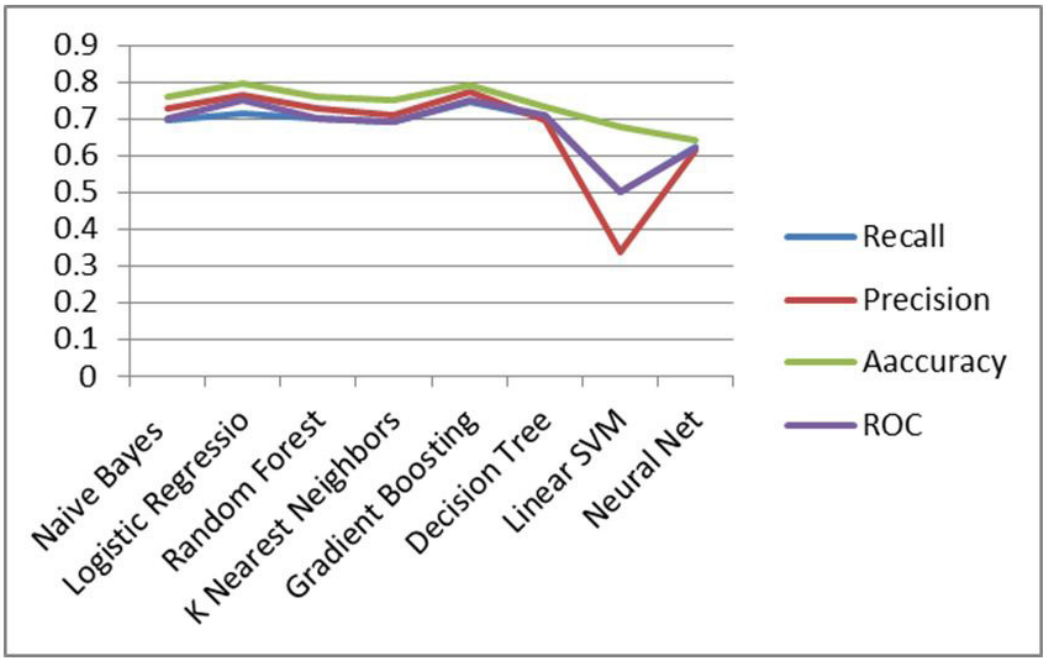


Figure 4: Comparison of other performance evaluation metrics of proposed and other algorithms

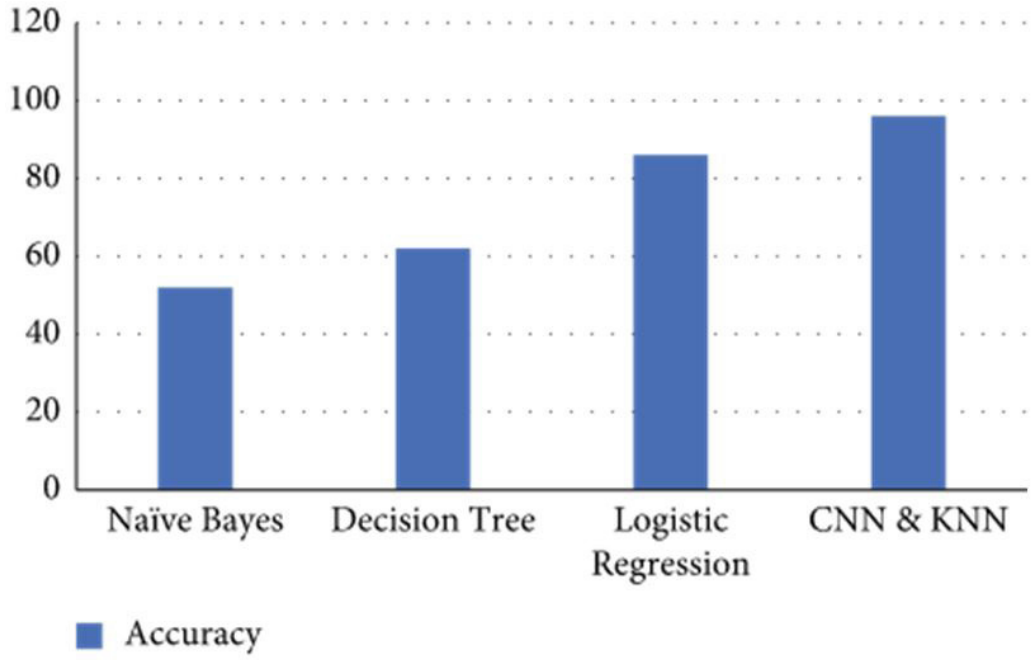


Figure 5: Shows a comparison of the suggested and other algorithms' accuracy

Conclusion

Using machine learning methods such as CNN and KNN, this research offered a method for detecting and predicting the existence of chronic illness in a person. The suggested method has the benefit of using both structured and unstructured data from real life to prepare data sets, which many existing techniques lack. The suggested model's performance is compared to that of other algorithms such as Nave Bayes, decision trees, and logistic regression methods in this research. The findings demonstrate that the suggested system has a 95 percent accuracy rate, which is greater than the other two methods.

The suggested method is thought to lower the risk of chronic diseases by detecting them sooner, as well as the cost of diagnosis, therapy, and medical consultations.

References

1. R. S. Akinbo, and O. A. Daramola, "Ensemble Machine Learning Algorithms for Prediction and Classification of Medical Images", in Machine Learning - Algorithms, Models and Applications. London, United Kingdom: IntechOpen, 2021.

2. L. Jena and R. Swain, "Work-in-Progress: Chronic Disease Risk Prediction Using Distributed Machine Learning Classifiers," 2017 International Conference on Information Technology (ICIT), 2017, pp. 170-173.

3. W. Yue, Z. Wang, J. Zhang and X. Liu, "An Overview of Recommendation Techniques and Their Applications in Healthcare," in IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 4, pp. 701-717, April 2021.

4. R. Ge, R. Zhang, and P Wang, "Prediction of chronic diseases with multi-label neural network," IEEE Access, vol. 8, pp. 138210–138216, 2020.

5. D. Gupta, S. Khare, and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," in Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 281–287, IEEE, Greater Noida, India, April 2016.

6. Manco, G., Ritacco, E., Rullo, A., Saccà, D., & Serra, E. (2022). Machine learning methods for generating high dimensional discrete datasets. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(2), e1450.

7. N. Peiffer-Smadja, T.M. Rawson, R. Ahmad, A. Buchard, P. Georgiou, F.-X. Lescure, G. Birgand, A.H. Holmes, in Machine learning for clinical decision support in infectious diseases: a narrative review of current applications, Clinical Microbiology and Infection, Volume 26, Issue 5,2020, 584-595.

8. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda published online ahead of print, 2022 Jan 13. J Ambient Intell Humaniz Comput. 2022;1-28.

9. P. Anandajayam, S. Aravindkumar, P. Arun and A. Ajith, "Prediction of Chronic Disease by Machine Learning," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1-6.

10. Aldhyani THH, Alshebami AS, Alzahrani MY. Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms. J Healthc Eng. 2020:4984967.

11. S. Ganiger and K. M. M. Rajashekharaiah, "Chronic diseases diagnosis using machine learning," in Proceedings of the 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), pp. 1–6, IEEE, Kottayam, India, December 2018.

12. G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," Journal of Personalized Medicine, vol. 10, no. 2, p. 21, 2020.

13. M. Gulhane and T. Sajana, "A Machine Learning based Model for Disease Prediction," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), 2021, pp. 1-5.

14. Gupta, Surbhi & Gupta, Manoj & Kumar, Rakesh. (2021). A Novel Multi-Neural Ensemble Approach for Cancer Diagnosis. Applied Artificial Intelligence. 1-36.

15. J. Peng, E. C. Jury, P. Donnes, and C. Ciurtin, "Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges," Frontiers in Pharmacology, vol. 12, p. 2667, 2021.

16. S. N. Induja and C. G. Raji, "Computational methods for predicting chronic disease in healthcare communities," in Proceedings of the 2019 International Conference on Data Science and Communication (IconDSC), pp. 1–6, IEEE, Bangalore, India, March 2019.

17. K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," Materials Today Proceedings, 2021.

18. K. Ng, A. Ghoting, S. R. Steinhubl, W. F. Stewart, M. Bradley, and J. Sun, "PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records," Journal of Biomedical Informatics, vol. 48, pp. 160–170, 2014.

19. Jose Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H. Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, Nathalie Conrad, Kazem Rahimi, Gholamreza Salimi-Khorshidi, In Deep learning for electronic health records: A comparative review of multiple deep neural architectures, Journal of Biomedical Informatics, Volume 101,2020,103337.

20. Caballé-Cervigón N, Castillo-Sequera JL, Gómez-Pulido JA, Gómez-Pulido JM, Polo-Luque ML. Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review. Applied Sciences. 2020; 10(15):5135.

21. M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Medical Informatics and Decision Making, vol. 11, no. 1, pp. 1–13, 2011.

22. S. Hegde and M. R. Mundada, "Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence-based feature selection approach," International Journal of Pervasive Computing and Communications, vol. 17, pp. 20–36, 2020.

23. Zhong, Hongye & Xiao, Jitian. (2017). Enhancing Health Risk Prediction with Deep Learning on Big Data and Revised Fusion Node Paradigm. Scientific Programming. 2017. 1-18.

24. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017.

25. A. I. Taloba, A. A. Sewisy, and Y. A. Dawood, "Accuracy enhancement scaling factor of viola- jones using genetic algorithms," in Proceedings of the 14th International Computer Engineering Conference (ICENCO), pp. 209–212, Cairo, Egypt, Decembe 2018.

26. Wang, Weilun & Chakraborty, Goutam & Chakraborty, Basabi. (2020). Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm. Applied Sciences. 11. 202.

27. Delpino, Felipe & Costa, Â.K. & Farias, S.R. & Filho, A.D.P. & Arcêncio, Ricardo & Nunes, Bruno. (2022). Machine learning for predicting chronic diseases: a systematic review. Public Health. 205. 14-25.

28. Maxwell, A., Li, R., Yang, B. et al. Deep learning architectures for multi-label classification of intelligent health risk prediction. BMC Bioinformatics 18, 523 (2017).

29. Ruiquan, Ge & Zhang, Renfeng & Wang, Pu. (2020). Prediction of Chronic Diseases with Multi-Label Neural Network.

30. X. Zhang, H. Zhao, S. Zhang, and R. Li, "A novel deep neural network model for multi-label chronic disease prediction," Frontiers in Genetics, vol. 10, p. 351, 2019.

CEOS
CENTER OF EXCELLENCE FOR OPEN SCIENCE

CEOS is an growing Open Access Publisher with diverse peer reviewed journals supported by Aca demic Editors, Scholars, Physicians, Instructors,

CEOS Publishers follow strict ethical standards for publication to ensure high quality scientific studies, credit for the research participants. Any ethical issues will be scrutinized carefully to maintain the integrity of literature.

CEOS Publishers believes scientific integrity and intellectual honesty are essential in all scholarly work. As an upcoming publisher, our commitment is to protect the integrity of the scholarly publica-tions, for which we take the necessary steps in all aspects of publishing ethics.

All the articles published in CEOS Publisher journals are licensed under Creative CommonsCC BY 4.0 license, means anyone can use, read and dow-nload the article for free. However, the authors reserve the copyright for the published manuscript.