Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230579

# Multi chronic disease prediction: A survey

**Chunduru Anilkumar[1, 2], Seepana Kanchana[1], Sasapu Bharath Kumar[1], Reddy Pravallika[1], Surapureddi Mrudula[1]**

[1]Dept of Information Technology, GMR Institute of Technology, Rajam, Andhra Pradesh-532127

[2]anilkumar.ch@gmrit.edu.in

**Abstract.** People today deal with a variety of illnesses as a result of their lifestyle choices and the environment. As a result, many people have chronic diseases that go untreated for long periods of time, imposing a tremendous impact on society. Therefore, predicting disease sooner is becoming a crucial duty. in order to systematically evaluate patients' future disease risks using their medical records. But for a doctor, making an accurate forecast based on symptoms is too challenging. The hardest task is making an accurate diagnosis of a condition. For this problem to be resolved, illness detection requires the use of deep learning and machine learning approaches. The amount of data in the medical sciences grows significantly every year. Earlier, health care for patient care has benefited from precise medical data analysis due of the development of information in the medical and healthcare areas. Prior identification and therapy are usually necessary to prevent chronic aeropathy from getting worse. Machine learning and deep learning algorithms are used to predict chronic diseases. Eight illness categories were our predictions. The Random Forest ensemble learning approach fared best overall. Finding the sickness prediction techniques with the highest accuracy and computation efficiency is the aim of this study.

**Keywords:** Machine Learning, Deep Learning, Convolutional Neural Network (CNN), Flask, Random Forest.

## 1. Introduction

Artificial Intelligence has given computers the ability to think more intelligently, which can Enable the computer to do things that were once impossible. AI researchers are studying machine learning as a subfield of their research. By making it possible to link medical events with the underlying causes of sickness, developments in computer and data storage technology have allowed the implementation of predictive machine learning methodologies in healthcare research. The longitudinal data on a patient's health that may be found in their electronic health records (EHR) opens up possibilities for predictive analysis. For instance, information from a series of clinical encounters involving hypertension may be connected to a later diagnosis of heart failure. There are many different types of machines learning techniques, including deep learning, evolutionary learning, reinforcement learning, and unsupervised, partially supervised, and supervised learning. These learning algorithms can quickly classify huge amounts of data. Patients who are unaware of the disease's course until a subsequent stage of diagnosis are a major contributor to the difficulty in forecasting chronic illnesses. We use a modeling technique

DOI: 10.54254/2755-2721/5/20230579

based on deep learning and machine learning to close this knowledge gap. Nearly five diseases, including diabetes, kidney, liver, heart, and breast cancer, are treated with the Random Forest algorithm in machine learning. On the basis of historical data, classification is done and decision trees are constructed. Therefore, the prediction algorithms are based on the decision tree when we provide the facts to be predicted. Given that the extra growth in the case of breast cancer is >4 cm in radius, the tree will immediately determine that the patient is already ill and in a dominating state. Medical data is also more complicated, high dimensional, heterogeneous, and irregular than normal data. While this was happening, interest in using deep learning to predict illnesses grew and produced a number of spectacular outcomes to increase prediction accuracy, strengthening the model. DNN has steadily created several network structures models to address a variety of medical data and challenges.

## 2. Related work

This essay is further broken into several pieces. The Literature Review section describes relevant studies on many of the diseases used in our study and the methodology of the algorithms used for prediction. They made disease predictions for three diseases he had, such as diabetes, stroke, and heart disease. [1]. In this post, we provide a method for leveraging the Flask API to forecast various illnesses. This article examines analyses of diabetes, diabetic retinopathy, breast cancer, and heart disease. Build model using a TensorFlow convolutional neural network and test it on the test set. The instance built achieved an accuracy of 91%. Here the disease is predicted based on user input the choice is given to the user [2]. Accuracy is increased when illnesses are predicted using a random forest with precise properties. By training 303 data instances, a random forest was created, and 10-fold cross-validation was used to verify correctness. His suggested model was discovered to be superior to other models after undergoing validation approaches, and by applying his 10-fold cross-validation, he was able to reach an accuracy of 85.81% [3]. Concerns of the author Chronic renal disease is defined as a brief decline in kidney function over months or years (CKD), often referred to as chronic renal disease. Early diagnosis of CKD aids in timely treatment of ill patients and also helps to stop the illness from growing worse [4].

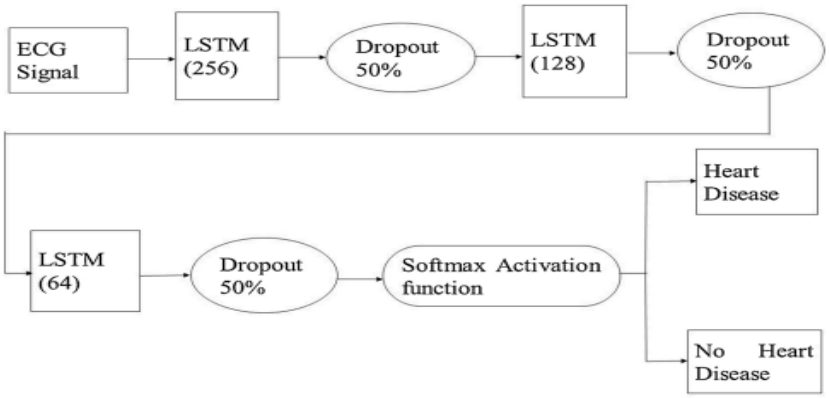


**Figure 1.** Flow chart of procedure.

Figure 1 explains how the computation is carried out till the expected result is reached. When the targeted class is evaluated to the estimated output. The resulting error is utilized to train the LSTM's system parameters design. The training process is continued by putting the training into practice. Sequential samples the training will continue till the defeat. Function achieves the lowest possible value Following successful training with the remaining 20%, the model's effectiveness is evaluated in phases. The suggested method uses the Health Care Cost and Utilization Project (HCUP) dataset to forecast a person's probability of developing an illness based on their medical diagnostic history. Published by the Healthcare Cost and Utilization Project (HCUP) for the purpose of developing disease prediction random forest classifiers. Without employing the sampling methodology, the second method contrasts the performance of RF, bagging, boosting, and SVM. [5] Diabetes, a disease in which blood sugar levels rise as a result of lack or low levels of insulin. They are using big data in healthcare to diagnose diabetes

DOI: 10.54254/2755-2721/5/20230579

disease. I used his WEKA tool for data analysis to make predictions [6]. In order to construct a multi-disease prediction, this research showed how big data and deep learning work together. Algorithms for hybrid deep learning are those. According to the research, the accuracy of the suggested JA-MVO-RNN + DBN was 2.8% better than any other hybrid approaches for diabetic condition. As a result, analysis has demonstrated the created hybrid deep learning's improved performance [7]. This study focuses on the use of operations on medical data produced in the field of medicine and health care to categories and forecast a specific condition. This white paper also outlines the techniques and types of data visualization techniques used in the project system and implemented on the medical data set to facilitate data understanding and data exploration [8]. Table 1 shows the comparison table of our research work which explains how different authors perform their methodology over disease prediction using different algorithms.

**Table 1.** Comparison table (continue).

| Author | Methodology/Description | Advantages |
|---|---|---|
| Dhiraj Dahiwade et.al (2019) | KNN,CNN | By providing patient data as input, which enables us to assess the amount of disease risk, An accurate general illness risk prediction was attained. |
| AkkemYaganteeswarudu et.al (2020) | Logistic regression, Random Forest, SVM. | The benefit of using a multi-disease prediction model in advance is that it can estimate the likelihood that numerous diseases. |
| Yeshvendra K. Singh et.al (2016) | Support vector machines, decision trees, random forests, logistic regression, and linear regression are some examples. | The Random Forest approach allows us to get the maximum accuracy possible. The suggested technique performs admirably on data from the actual world. |
| Sai VaishnaviAvilala et.al (2019) | KNN, Random forest, NaiveBayes. | The work of information mining is used to the declining value of enormous knowledge storage systems and the ability to apply computationally. |
| Mohammed Khalilia1 et.al (2011) | SVM, Random Forest, Bagging, Boosting | Able to overcome the challenge of class inequality and achieve desirable results. |
| Arwatki Chen Lyngdoh et.al (2021) | KNN,Naivebayes, Decision tree classification , Random forest, SVM | The results achieved proved the adequacy of the system with anaccuracy of 76% using KNN. |
| AnushaAmpavathi et.al (2021) | Deep belief Network(DBN) and Recurrent Neural Networks (RNN) | Deep neural network hybrid technique (RNN+DBN) outperforms all other approaches with a high accuracy of 2.8%. |

DOI: 10.54254/2755-2721/5/20230579

**Table 1.** (continued).

| | | |
|---|---|---|
| AjinkyaKunjir et.al (2017) | Genetic Algorithm, J48 Decision Tree, Naive Bayes. | The methods and varieties of data visualization approaches are described in this study to make it simple to comprehend and explore data. |
| Tingyan Wang et.al (2019) | RNN with LSTM | According to the study, recurrent neural networks with LSTM units function effectively at various degrees of diagnosis aggregation. |
| AdyashaRath et.al (2021) | GAN with LSTM | Five models are tested and out of them, GAN with LSTM model performs well where the Naive Bayes performs less. |

In order to suit the demands of many stakeholders, medical diagnoses based on the International Classification of Diseases (ICD) are aggregated into several levels for prediction. When utilizing 3-digit ICD code aggregation, LSTM networks provide equivalent scores of 98.90% in the MIMIC dataset and 95.12% in the GenCare dataset, respectively, while achieving 96.60% and 96.83% when using four-digit ICD code aggregation for these two datasets [9].

**Table 2.** Training and testing computational time.

| Classifier | Training Time | Testing Time | Computational Time |
|---|---|---|---|
| KNN | 2.7549004457 | 5.5154268899 | 7.90916633 |
| NB | 0.18233000002 | 0.0567890879 | 0.23911908 |
| DT | 0.77532997643 | 0.0478995599 | 0.82322953 |
| RF | 12.6574437890 | 0.9546667586 | 13.6121105 |
| SVM | 670.49820000 | 15.954679321 | 686.452876 |

Table 2 shows that the SVM classifier has a positive influence. It consumes the majority of processing time and is highly expensive in terms of processor and memory utilization. Furthermore, it is not providing greater accuracy than other KNN. the SVM classifier chooses the greatest computing time and is hampered by model overfitting.

This paper describes the classification of heart disease from electrocardiogram samples using deep learning techniques. With an F1 score and an AUC of 0.993, 0.994, and 0.995 for the PTB ECG dataset, the GAN-LSTM model outperforms every other model. [10]

## 3. Methodology

Multi-Disease Prediction is the capacity to simultaneously detect various illnesses that a person is anticipated to develop throughout time [11]. To anticipate the occurrence of chronic disorders in people, the approach proposed here employs two categorization algorithms. Unit CNN and Random Forest classifiers are the classifiers in use. Figure 2 shows the Architecture of Our project's which major goal is to offer a prediction system that uses the patient's test report findings as input to determine whether or not the patient will experience the disease.Nginx is a web server that may also be utilized as a reverse proxy, load balancer, uWSGI is an application of the WSGI specification that outlines how a web server can connect with a web application. The system then displays the prognosis through an interface.

There are 3 steps to it:

a) Data Pre-processing

b) Using CNN and Random Forest for classification

DOI: 10.54254/2755-2721/5/20230579

c)Presenting the outcome with Flask Web frame

Patients' records may have statistical data and images in form of x-rays and so on .our model is compatible to work with any type of data for classification. random forest takes statistical data for classification and CNN use images for classification. An effective machine learning algorithm the supervised learning approach includes Random Forest. It may be used to solve classification and regression-related ML problems. It draws on the notion of ensemble learning, which is a technique for combining many classifiers to solve challenging issues and enhance model performance. As its name suggests, Random Forest is a classifier that averages several decision trees applied to various subsets of the supplied information to improve the predicted accuracy of the dataset. Instead of depending exclusively on one decision tree, the random forest takes predictions from all decision trees and forecasts the result based on the votes of the majority of predictions.

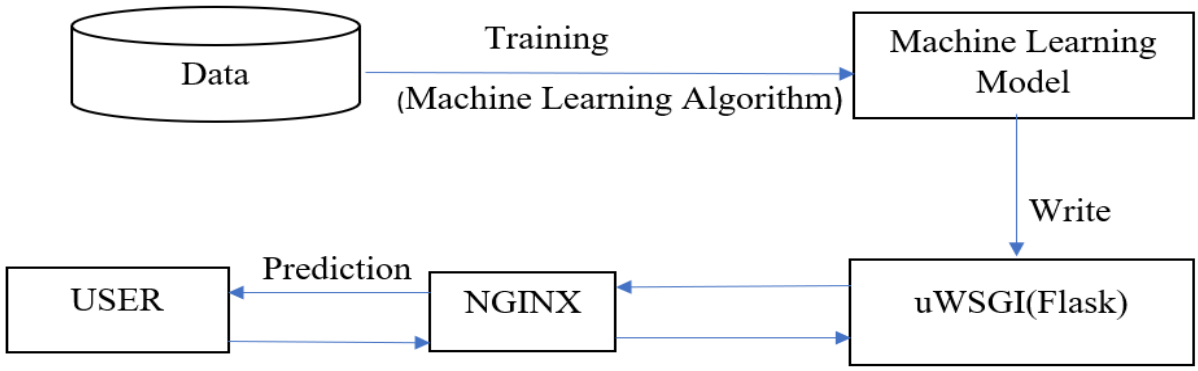


**Figure 2.** Architecture of procedure.

Step-1: Pick K data points at random from the practise set.
Step-2: Construct a decision tree using the chosen data points (subset).
Step-3: Decide how many N-tree decision chains to construct.

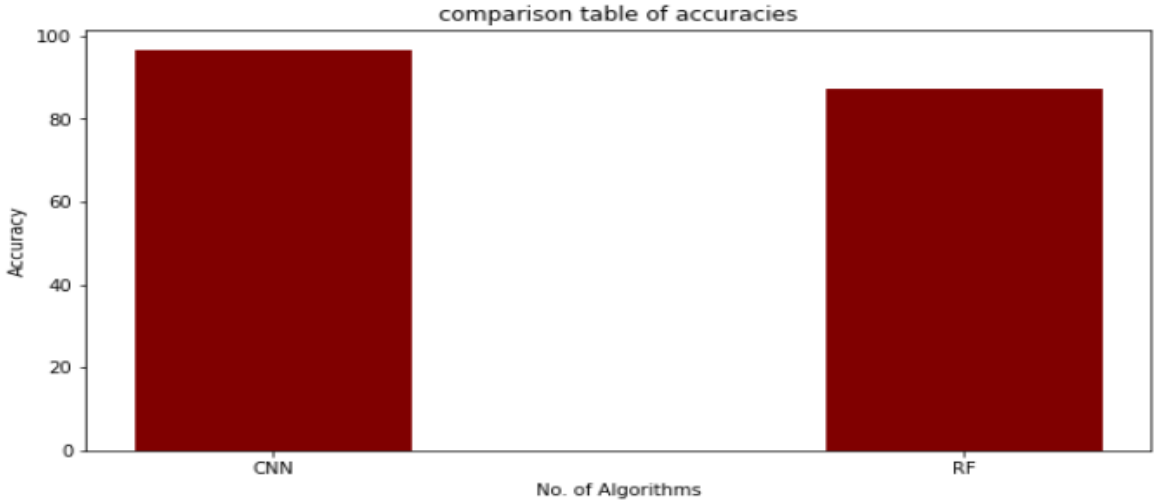


**Figure 3.** Comparison of algorithms.

The Figure 3 shows the comparison graph of CNN and Random Forest. We apply both algorithms on the same Malaria dataset which shows that the CNN algorithm performs well on images of malaria disease detection compared to the Random Forest.

CNN receives a picture as input, which is then categorised and processed. Convolutional layers, pooling, fully connected layers, and filters are applied to each input picture in a series in CNN (Also known as kernels). After then, the Soft-max function will be used to determine whether or not a person

DOI: 10.54254/2755-2721/5/20230579

has a sickness. By using above classifiers, we predict 7 types of diseases and Displaying result using Flask Framework.

## 4. Conclusion

In order to forecast multiple diseases, CNN and Random Forest are used in this study. We can contrast the effectiveness of the currently being utilized classifiers with that of other classifiers in use. Early detection aids in the timely treatment of those who are ill and also helps stop the illness from getting worse. The need for the medical zone is for early diagnosis and prompt treatment of the illness.

## References

[1] Dahiwade, Dhiraj, GajananPatle, and EktaaMeshram. "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, 2019.

[2] Yaganteeswarudu, Akkem. "Multi disease prediction model by using machine learning and Flask API." In 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 1242-1246. IEEE, 2020.

[3] Singh, Y.K., Sinha, N. and Singh, S.K., 2016, November. Heart disease prediction system using random forest. In International Conference on Advances in Computing and Data Sciences (pp. 613-623). Springer, Singapore.

[4] Devika, R., Sai VaishnaviAvilala, and V. Subramaniyaswamy. "Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest." *2019 3rd* International conference on computing methodologies and communication (ICCMC). IEEE, 2019.

[5] Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. "Predicting disease risks from highly imbalanced data using random forest." BMC medical informatics and decision making 11.1 (2011): 1-13.

[6] Lyngdoh, Arwatki Chen, Nurul Amin Choudhury, and SoumenMoulik. "Diabetes Disease Prediction Using Machine Learning Algorithms." In 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 517-521. IEEE, 2021.

[7] Ampavathi, Anusha, and T. VijayaSaradhi. "Multi disease-prediction framework using hybrid deep learning: an optimal prediction model." Computer Methods in Biomechanics and Biomedical Engineering 24.10 (2021): 1146-1168.

[8] Kunjir, Ajinkya, HarshalSawant, and Nuzhat F. Shaikh. "Data mining and visualization for prediction of multiple diseases in healthcare." 2017 International Conference on Big Data Analytics and Computational Intelligence *(ICBDAC)*. IEEE, 2017.

[9] Wang, Tingyan, Yuanxin Tian, and Robin G. Qiu. "Long short-term memory recurrent neural networks for multiple diseases risk prediction by leveraging longitudinal medical records." IEEE journal of biomedical and health informatics 24, no. 8 (2019): 2337-2346.

[10] Rath, Adyasha, Debahuti Mishra, Ganapati Panda, and Suresh Chandra Satapathy. "Heart disease detection using deep learning methods from imbalanced ECG samples." Biomedical Signal Processing and Control 68 (2021): 102820.

[11] Subramanian, M., Lv, N. P., & VE, S. (2022). Hyperparameter optimization for transfer learning of VGG16 for disease identification in corn leaves using Bayesian optimization. Big Data, 10(3), 215-229.