International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning Approach for Early Disease Prediction and Risk Analysis

Rutuja A Gulhane[1], Sunil R Gupta[2]

Submitted: 14/09/2022 **Accepted:** 23/12/2022

Abstract:

The optimistic future of AI integration in healthcare has become more imaginable in recent years because to AI's rapid development and the steady launch of AI research in the medical profession. The most significant promise for machine learning has been in applications like predicting how well a given drug will work. Clinicians have significant challenges when manually diagnosing abnormalities; this study aims to detect and predict people suffering from many diseases. Over time, both the accuracy and clarity of the pharmaceutical illness forecast have increased from the initial logistic regression to the machine learning prototype. This article takes a look at the many different machine learning frameworks available, as well as several common diseases and a brief explanation of the machine learning prediction methods used for each. Find the flaws in the current illness projection and estimate its growth going forward. Its overarching goal is to demonstrate ML's utility for disease prediction and to highlight the vital connection between ML and emerging medical technologies. The various feature extraction strategies available inside machine learning technologies may maintain their relevance in the field of medical study in the years to come.

Keywords: Precision in Disease Prediction; AI, ML.

1. Introduction

The adoption of IT in healthcare has accelerated its development in recent years [1]. Similar to how cell phones have simplified people's lives, the integration of IT into healthcare makes people's lives more cost-effective and comfortable [2]. To this end, it may be possible to implement forms of artificial intelligence into healthcare, such as the creation of smart ambulances and hospitals that provide a range of helpful services to both patients and doctors [3, 4]. Results obtained from a combination of structured and unstructured data are much more accurate than those obtained from structured data alone. Combining structured data like affected person demographics, and complaints can be helpful [5, 6]. Difficulty in diagnosis is a common problem with rare disorders. For this reason, distinguishing between patients with rare diseases and those with more common chronic ailments is facilitated by the use of self-reported interactive information. It is believed that the detection of rare diseases can be greatly improved by combining surveys with machine learning techniques [7].

Statistics from magnetic resonance imaging (MRI), ultrasonography, data from social media, and electronic recordings of movement, behaviour, and experimentation have all been produced in the past decade to speed up the collection of data. Due to the multidimensional nature of healthcare large data sets, the total number of observations may exceed the total number of attributes recorded for each observation [8]. The applications of machine learning are expanding rapidly. As the field of health care epidemiology nears a watershed point, more extensive training data benefits several challenging models [9]. This data has the potential to enhance our knowledge of disease risk features, leading to better measures of infection control in healthcare facilities, risk stratification for patients, and the identification of the mechanisms by which infectious illnesses are spread [10]. By analysing test results and other patient data, machine learning can aid in the early detection of illness. By searching for relevant information in the database, it may be possible to generate high-level knowledge that sheds light on sickness patterns and enables earlier diagnosis [11]. Preprocessing the data used to construct an information set for missing values and selecting only the most relevant variables for an effective disease prognosis are two ways to boost prediction accuracy and speed up model training [12].

[1]*Research Scholar, PRMIT&R, Badnera, Maharashtra, India*
gulhanerutuja@gmail.com
[2]*Assistant professor, PRMIT&R, Badnera, Maharashtra, India*
sunilguptacse@gmail.com

Most people born after the advent of the Internet and modern technologies care little about their physical well-being. Nobody goes to the hospital for regular checkups anymore because they're too preoccupied with their smartphones. To take advantage of this trend, we need to develop a machine learning model that can identify the severity of a patient's symptoms and determine the possibility that they have a disease or may develop one in the near future [13, 14]. Early detection increases the likelihood of a successful treatment outcome, whereas prevention is possible thanks to early detection [15]. Some examples of machine learning techniques include supervised learning, semi-supervised learning, unconfirmed learning, learning by strengthening, learning by evolution, and deep learning. The difficulty comes from trying to process vectorized characteristics derived from real data [16]. It is the ideal combination of these vectors that determines how well a process works. However, in most cases, this is not possible due to the great complexity of the trajectories or information disparities. Therefore, it is crucial to construct an information set with a highly suitable dimension, even if doing so requires reducing the dimensionality of the information set. An increase in model accuracy is observed after this dimensionality reduction [17].

Patients with these diseases must have access to accurate medical diagnosis and treatment information, making the illness management plan essential [18]. An improved method of patient self-management [19] is the use of mobile apps that gather patient health information. It is impossible to make a reliable prognosis for a condition without the patient's symptom history, information about their visits to experts, laboratory findings, computed tomography, and X-ray pictures [20]. There is a lack of studies evaluating the accuracy and predictability of machine learning models for illness identification based solely on laboratory test findings. Moreover, the quality of presentations can be enhanced through the application of collaborative machine learning and deep learning models [21, 22]. The coordinating the best time for medical practitioners to complete a wide variety of tasks that cannot be mechanized, are two examples of the significant roles artificial intelligence (AI) plays in the health care industry [23].

The proposed method relies on machine learning to identify and foresee human health problems. The data set includes both structured information like the patient's age, gender, height, weight, etc. (but not personally identifiable information like their name or ID) and unstructured information like the patient's symptoms, doctor-patient conversations, and lifestyle choices related to their illness. Data is preprocessed to recover hidden values, and then reconstructed to enhance model quality and prediction precision. Machine learning techniques like Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and the Planned approach (HDO) are used for making predictions.

The following is the structure of the study: The research methodology, including sections describing the associated works authorised out throughout the investigation, the processes utilised, the results and discussion, the conclusion, and the list of sources used, are all detailed in this report's Appendices.

2. Related Work

This section describes the efforts put into creating the prototype suggested for predicting various illnesses. A literature review was conducted to help in the creation of the proposed system, and this information will be presented in the following presentations.

The SVM algorithm and the Naive Bayes approach [11] were used to make predictions about renal illness. Using the common ANFIS method, the authors made an effort to categorise renal illness into its several stages. The investigation set out to develop a classification system that performs well across a number of important metrics, including accuracy and runtime. The Naive Bayes Algorithm was faster and more accurate than the SVM Procedure for classification. The results demonstrate that SVM performs better than the Naive Bayes Approach in terms of predicting renal disease.

Predictions of cardiac issues were made using a fuzzy method that relied on a membership function [12]. To eliminate confusion, the authors implemented the Fuzzy KNN Classifier. There were a total of 550 records in the dataset, which were divided into 25 categories.

Using the ANN algorithm, a novel method for identifying heart illness at an early stage has been developed [13]. The authors of a recent paper [14] developed a computer approach for resolving complex concerns about the prognosis of cardiac disease. This clever system, developed using the Naive Bayes method, is superior in speed, accuracy, and overall quality. It might aid in experimentally judging cardiac spells, which could help clinicians. Possible enhancements to this setup include integrating SMS support, developing native apps for Android and iOS, and installing a pacemaker.

Together, the flexibility of support vector machines and other factors led to the discovery of diabetes and breast cancer [15]. The idea behind adaptive SVM was to provide a diagnostic method that could be used quickly, automatically, and in a variety of different ways. Traditional SVM had its bias value adjusted for enhanced performance. The proposed classifier delivered the result as a series of "if-then" clauses. Both diabetes and breast cancer were detected with perfect precision using the proposed strategy[16].

[17] developed a unique strategy for identifying cardiovascular disease risk variables using machine learning procedures in 2019, which resulted in enhanced prediction precision. The test analysis was used by the hybrid model to detect the properties of machine learning techniques.

In 2019, Xiao et al. [18] published a novel integrated technique called MRLDC for predicting illnesses linked with circRNA entrants.

2.1 Clinical Application Issues: In the present day, most systems that attempt to merge deep learning with disease prognosis are still in the ideation phase and have not been put through any kind of clinical testing. Several potential causes contribute to the current state of affairs.

• Security

• Confidentiality

2.2 Problem Statement: According to studies, most machine learning models developed for healthcare analysis only consider one disease. The first is for analysing liver issues, the second is for analysing cancer, and the third is for analysing lung issues. In order to make accurate predictions about a wide range of illnesses, a person will need to consult a number of different online resources. The predicting of several diseases using a single analysis is not a well-defined process. Poor accuracy in some models has major consequences for patient health. Companies that want to assess their patients' medical records face increased time and financial investment due to the need to install various models. Some current systems take into account too few parameters, which might lead to inaccurate outcomes..

3. Methodology

The purpose of this article is to create a strong machine learning model capable of reliably forecasting a human's status based on symptoms.

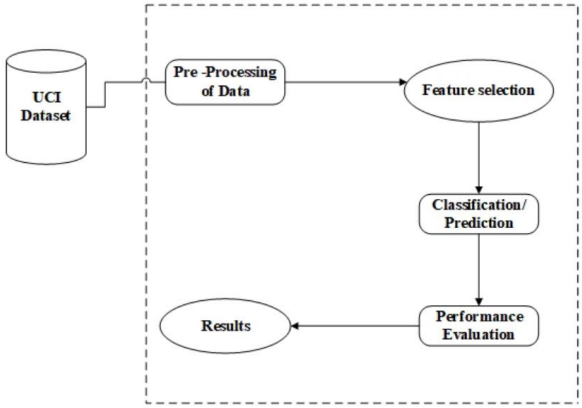

Fig 1. Process of Evaluation

Machine learning was used to achieve this result.

Collecting relevant data is the first step in any machine learning project. This endeavour makes use of data from the Kaggle dataset. This test and training set has images and CSV files. The dataset lacks symptom columns, which would allow for their representation.

• Data Cleansing: Data cleaning is the most crucial part of any machine learning programme. It is the quality of our data that will define how well our machine learning model performs. This means that data cleansing is an essential step before training any model. Supports in our dataset are all numbers, however the objective column, prediction, is text that was encoded using a labeler to become a number.

After data is collected and cleaned, it can be used for training a machine learning model in the next step, "model construction." All three of these classifiers—Support Vector, Naive Bayes, and Random Forest—will be trained using this cleansed data. The models' accuracy will be evaluated using a confusion matrix.

After training the three representations, inferred disease for input indications is predicted by combining predictions from the three representations. Consequently, the total reliability and precision of our projection will increase.

At last, a function has been developed that accepts a list of symptoms separated by commas as input, predicts the illness using the appropriate models, and outputs the results as JSON.

3.1 Model Building

Data partitioning is the first step in the modeling process. K-Fold cross-validation was used to check the ML models' representations. The cross-validation procedure used three different classifiers: the Support Vector Classifier, the Gaussian Naive Bayes Classifier, and the Random Forest Classifier.

3.2 Proposed Approach

This paper proposes a hybrid meta-heuristic approach, the HDO, by fusing the Red Deer Algorithm (RDA) and the Divide and Conquer (DA) algorithm (Dragonfly Algorithm). This HDO uses the Efficient Recurrent Neural Network (E-RNN) model, which optimises the "number of hidden neurons" and the number of epochs of a recurrent neural network. It guarantees a high detection rate for cardiac and breast cancer diagnostics [20] by maximising performance through accuracy and precision.

We present an alternate HDO algorithm that incorporates aspects of RDA and DA techniques. The researchers decided for RDA because it strikes a better balance between the exploration and exploitation stages, has a higher rate of convergence, and gets to global optimal solutions faster. However, RDA's multiple

"controlling factors," poor execution rates, and inability to define global solutions make it a challenge to optimize [21]. In order to improve the cohesion of the phases, increase the number of viable solutions, and strengthen the algorithm as a whole, it is motivating to create a new HDO algorithm that draws inspiration from the DA approach. In light of the loosened rules, this is now much simpler to accomplish. The goal of these hybrid DA/RDA approaches is to obtain optimal solutions at a speedy convergence [22].

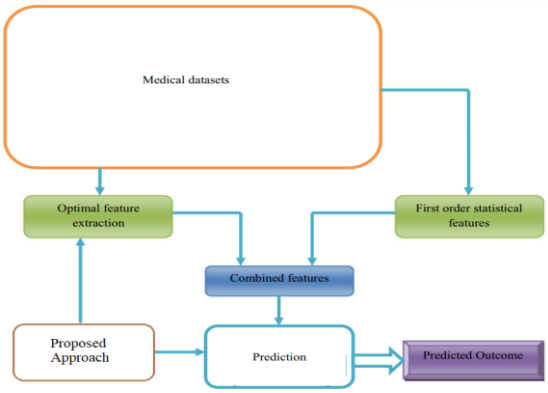

Fig 2. Flow of Proposed Approach

3.3 Performance Evaluation Criteria

The suggested system's efficacy is evaluated using several metrics, such as accuracy, precision, recall, and F1 measures. The term "true positive" is used to denote the amount of events that correspond to the requirements that were originally predicted. A false positive is an overestimation of the frequency with which some event is expected to occur. The term "true negative" is used to describe the number of events that were correctly predicted to be optional. The number of events that were incorrectly predicted as optional is known as the "false negative." Possible explanation for how the four measuring modalities came to be:

This is an accurate statement. This indicates how closely an extent approximates the true value, and it can be calculated with the use of the formula given in the following section.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}} \quad (1)$$

- Using the method below, which takes into account the consistency of data from several measurements, you can determine the precision of your measurements.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

The fraction of relevant outcomes that the model correctly categorizes can be calculated using the real formula.

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

Maintaining an unwavering commitment to accuracy can, on occasion, lead to misleading conclusions. There are times when using a model with a lower level of accuracy might be advantageous since it provides a more reliable prognosis for the problem [13]. When there is a considerable class imbalance in the issue domain, it is possible for all forecasts to be made using the value associated with the dominating class. As a direct result of this, we favor utilizing the four separate variables so that we can obtain more specific results.

4. Result Analysis & Discussion

4.1 Experimental setup

On a machine running Windows 10 and equipped with 8 gigabytes of RAM and an Intel i3 processor, the proposed method is implemented using Python. During the course of the experiment, a population size of 10 was thought to exist, and a maximum of 10 iterations were investigated. The proposed HDO's performance evaluation was compared to that of four different meta-heuristic algorithms: CNN, KNN, DT, and SVM. The phrases "accuracy, sensitivity, specificity, and precision, as well as the F1 score," were among the evaluation criteria.

4.2 Description of Datasets

When analysing the likelihood of various diseases, databases like those pertaining to breast cancer, heart disease, lung disease, and thyroid conditions are taken into account.

Cancer of the breast "From a digitised picture of a fine needle aspirate of breast mass, features are generated that identify the properties of cell nuclei in the image," the author writes. "This allows for a more accurate diagnosis of breast cancer."

In the case of heart disease, every published test only addressed a subset of the condition's 76 symptoms. Particularly, researchers keep turning to the Cleveland dataset for their work.

"Hong and Young employ it to demonstrate the efficiency of the optimum discriminant plane in ill-posed settings, specifically with regard to lung cancer. The findings are broken down into three distinct groups based on the histological characteristics of the lung tumours. It is made up of characteristics such as class and other nominal predictive traits, taking into account numeral values that range from 0 to 3.

The final weight, age, marital status, work class, education-number, occupation, association, native

country, sex, capital damage, teaching, hours per week, race, and financial gain are all factors that are considered while assessing thyroid disease.

4.3 Performance Evaluation

The performance estimation of the designed and proposed approach-assisted "multi illness diagnosis model" is displayed in the table below. This table compares this model's performance to that of existing algorithms. Based on the findings of the evaluation, the accuracy of the designed and proposed HDO model has shown their efficiency to be 3.7%, 3.8%, 1.5%, and 1.8% more advanced than SVM, CNN, KNN, and DT, respectively. When comparing the designed model to already existing algorithms, this ensures that the designed model has greater effectiveness.

Table I: Evaluation on the Proposed Approach over Conventional Algorithms

TERMS	SVM	CNN	KNN	DT	Proposed (HDO)
"Accuracy"	94.22581	89.56452	86.48387	88.40323	95.72581
"Sensitivity"	94.11953	89.6133	86.25759	88.30297	95.74944
"Specificity"	94.33409	89.51482	86.71443	88.50537	95.70173
"Precision"	94.42129	89.6993	86.86836	88.67137	95.78005
"F1-Score"	94.27017	89.65628	86.5619	88.48679	95.76474

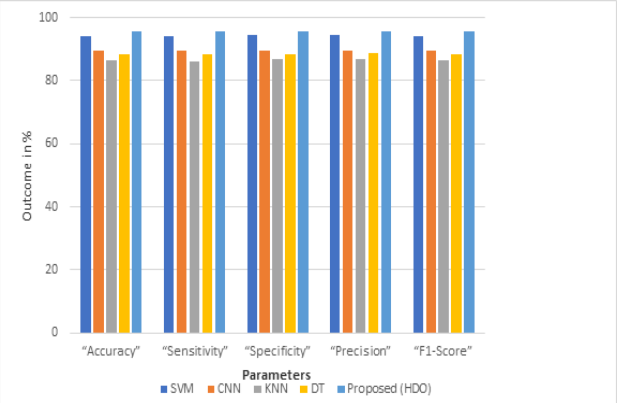

Fig 3. Performance Analysis of Proposed Method with Existing Approaches

5. Conclusion

There is little doubt that the dawn of the digital age and the Internet of Things will herald in a time when medical treatment can better use exciting new possibilities. As Well-known machine learning prediction algorithms and frameworks for various diseases are introduced in this article, along with the optimal selection strategy based on commonly-used medical diagnostic data. Also covered are the many diseases that can be anticipated using these techniques. Many of the current medical industry's practical difficulties and deep learning's inherent limits have been discovered. It is proposed that precision medicine and basic research into individual diseases should work together, that the Internet of Things and the creation of intellectual medical stages and equipment should work hand in hand, and that model development ideas be taken into account in light of the uniqueness of medicinal data. We predict that in the not-too-distant future, both illness analysis and machine learning will diversify to the point where they can identify illnesses with decision-making abilities beyond those currently possible. Multiple illness types have seen an uptick in the number of ML models created to study them. Having interconnected representations that can teach one another makes the ML system more robust. This relates to the complex organization of medicine, which will help advance diagnostics and practical applications of medicine, ultimately leading to a larger medical community.

6. Future Scope

For the sake of future expansion, we plan to include additional diseases in the current system.

• Make the system as user-friendly as possible and offer a chatbot to handle common questions.

References:

[1]. R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Curr. J. Appl. Sci. Technol. 40 (6) (2021) 78–89, https:// doi.org/10.9734/cjast/2021/v40i631320.

[2]. M. Sivakami, P. Prabhu. Classification of algorithms supported factual knowledge recovery from cardiac data set, Int. J. Curr. Res. Rev. 13 (6) 161- 166. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).

[3]. M. Sivakami, P. Prabhu. A Comparative Review of Recent Data Mining Techniques for Prediction of Cardiovascular Disease from Electronic Health Records. In: Hemanth D., Shakya S., Baig Z. (eds) Intelligent Data Communication Technologies and Internet of Things. ICICI (2019).

[4]. P. Prabhu, S. Selvabharathi. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus. In 2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC (2019), 138–142, Institute of Electrical and Electronics Engineers Inc. ISBN:9781728136639. 2019.

[5]. N. Jothi, N.A. Rashid, W. Husain, Data mining in healthcare – A review, Procedia Comput. Sci. 72 (2015) 306–313.

[6]. H. Polat, H. Danaei Mehr, A. Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, J. Med. Syst. 41(4) (2017).

[7]. K.B. Wagholikar, V. Sundararajan, A.W. Deshpande, Modeling paradigms for medical diagnostic decision support: a survey and future directions, J. Med. Syst. 36 (5) (2012) 3029–3049.

[8]. E. Gürbüz, E. Kılıç, A new adaptive support vector machine for diagnosis of diseases, Expert Syst. 31 (5) (2014) 389–397.

[9]. M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification, Expert Syst. Appl. 41 (5) (2014) 2239–2249.

[10]. Y. Kazemi, S.A. Mirroshandel, A novel method for predicting kidney stone type using ensemble learning, Artif. Intell. Med. 84 (2018) 117–126.

[11]. H. Barakat, P. Andrew, Bradley, H. Mohammed Nabil Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, IEEE Trans. Inf. Technol. Bio Med. J. 14 (4) (2009) 1–7.

[12]. R. Tina Patil, S.S. Sherekar, Performance analysis of Naive bayes and J48 classification algorithm for data classification, Int. J. Comput. Sci. Appl. 6 (2) (2013) 256–261.

[13]. Shruti Ratnakar, K. Rajeswari, Rose Jacob, Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes, Int. J. Adv. Comput. Eng. Netw. 1 (2) (2013) 51–55.

[14]. S. Grampurohit, C. Sagarnal, Disease prediction using machine learning algorithms, Int. Conf. Emerg. Technol. (INCET) (2020) 1–7, https://doi. org/10.1109/INCET49848.2020.9154130.

[15]. R.J.P. Princy, S. Parthasarathy, P.S. Hency Jose, A. Raj Lakshminarayanan, S. Jeganathan, Prediction of Cardiac Disease using Supervised Machine Learning Algorithms, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570–575, https://doi.org/10.1109/ICICCS48265.2020.9121169.

[16]. P. Deepika, S. Sasikala. Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization, 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1068-1072, doi: 10.1109/ ICECA49313.2020.9297398.

[17]. Haq A U, Li J P, Memon M H, Nazir S and Sun R 2018 A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms; Mob. Inf. Syst. 2018: 21–22

[18]. Xiao Q, Luo J and Dai J 2019 Computational prediction of human disease- associated circRNAs based on manifold regularization learning framework; IEEE J. Biomed. Health. 23 2661–2669

[19]. Yang X, Lu R, Shao J, Tang X and Yang H 2019 An efficient and privacy-preserving disease risk prediction scheme for E-healthcare; IEEE Internet Things. 6 3284–3297

[20]. Va´squez-Morales G R, Martı´nez-Monterrubio S M, MorenoGer P and Recio-Garcı´a J A 2019 Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning; IEEE Access. 7 152900–152910

[21]. Minhas S, Khanum A, Riaz F, Khan S A and Alvi A 2018 Predicting progression from mild cognitive impairment to Alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers; IEEE J. Biomed. Health. (2020) 818–825.

[22]. Escudero J, Ifeachor E, Zajicek J P, Green C, Shearer J and Pearson S 2013 Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease; IEEE T Bio-Med Eng. 60 164–168.

[23]. Haq A Q, Li J P, Memon M H, Khan J, Malik A, Ahmad T, Ali A, Nazir S, Ahad I and Shahid M 2019 Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings; IEEE Access 7: 37718–37734.

[24]. Sierra-Sosa D, Garcia-Zapirain M B, Castillo C, Oleagordia I, Nun˜o-Solinis R, Urtaran-Laresgoiti M, Elmaghraby A 2019 Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs; IEEE T. Ind. Inform. 15: 5682–5689.