Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P.2790-2802 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Hybrid Machine Learning Model for Chronical Disease Prediction

[1] Suman, [2]Yudhvir Singh, [3]Neha Gulati

[1]Research Scholar, C.S.E. UIET Department ,MD University, Rohtak
[2] Professor C.S.E. UIET Department, MD University, Rohtak
[3]Assistant Professor, Punjab University, Chandigarh
*punia.suman@gmail.com

How to cite this article: Suman, Yudhvir Singh, Neha Gulati (2024) Hybrid Machine Learning Model for Chronical Disease Prediction. *Library Progress International*, 44(3), 2790-2802.

Abstract
The existing methods for predicting chronic diseases have a number of issues with feature extraction, outlier removal, and classification. The primary goal of this research project is to address current shortcomings. The prediction of chronic diseases will be more accurate once the current flaws are eliminated. As a result, it is critical to analyse the changes in chronic diseases using epidemiological data and to forecast the disease using socioeconomic data. In this study, a hybrid model based on CNN and Bidirectional LSTM is given. For feature extraction, CNN is utilized, while logistic regression is employed for classification. Using CNN, the collected features were fed into a bidirectional LSTM for classification. The suggested framework is put into practice using Python, and the accuracy, precision, and recall of the output are examined.

Keywords: Chronical Disease, Machine Learning, Hybrid Model, CNN, Bidirectional LSTM

Introduction

NCDs (Noncommunicable diseases), sometimes referred to as chronic diseases, are distinguished by their protracted duration and slow rate of development. Chronic respiratory disorders, diabetes, and cardiovascular diseases (such as hypertension, coronary artery disease, and stroke) are common forms of chronic illness. The extended nature of these illnesses and the substantial costs associated with their medical treatment impose significant economic strains on both society and individual families [1][2]. Currently, the global effort to prevent and manage chronic diseases remains challenging, particularly in low- and middle-income countries. These diseases contribute significantly to the overall disease burden, and the number of deaths they cause is steadily rising. Chronic diseases, with their complex causes, often lead to complications that make it challenging for doctors to find suitable treatments. Despite the availability of advanced medical technologies and devices, such as wearable health monitors and mobile health applications, which can continuously track and gather extensive patient data, this information is often heterogeneous, encompassing laboratory test results, physical measurements, and electrocardiograms. Physicians leverage methods such as machine learning and data mining to understand and comprehend this complicated data, assisting them in making the best treatment decisions by summarizing and analyzing the information. [3][4]. One subset of AI (artificial intelligence) approaches is ML (machine learning). AI is a technology that simulates intelligent behavior with minimum human intervention by using computer information. This concept is widely believed to have originated with the advent of robotics. Due to the rapid advancements in electronic speeds and programming, computers may soon demonstrate intelligent behaviour comparable to that of humans. These capabilities are driven by significant progress in contemporary AI theories. Essentially, artificial intelligence is human-like intelligence carried out using devices. In computing technology, AI is characterized by a machine's ability to independently emulate intelligent behaviour using ML [5][6]. The application of AI (Artificial Intelligence) in medicine is rapidly advancing. AI in medicine involves the use of automated diagnostic procedures and patient care. The increased adoption of AI in diagnosis and treatment will automate a significant portion of these tasks, allowing medical professionals to focus more on non-automatable aspects of patient care. As a result, AI technology is increasingly being utilized in various domains, including HR (human resources).

1.1. Research Gaps

The chronic disease prediction dataset is the complex dataset due to which it is difficult to establish relationship between various attributes. When proper relationship is not established between the attributes then it is difficult to identify main characteristics for predicting of chronic illness. In the previous research work, no technique of outlier remove is implemented for the efficient prediction. The outlier removal will clean and also categorize the dataset for the further processing. In the pre-processing phase standard technique needs to propose like under sampling or over sampling to clean the dataset. When the dataset is cleaned properly then it leads to efficient prediction analysis [9]. This study focuses on data mining predictions for chronic diseases. Three fundamental processes comprise the chronic illness prediction technique: pre-processing, extracting features, and classification. In the pre-processing stage, excess and missing values are removed from the data set. The association between the target set and the characteristic will be established in the following stage. The entire dataset will be split into a training and test set during the final stage [10]. We will analyze the ensemble model's effectiveness with regard to of F1 measure, recall, sensitivity, and precision. In order to forecast chronic diseases, the ensemble classifier's high complexity must be decreased.

1.2. Need of Study

This study focuses on data mining predictions for chronic diseases. Three fundamental processes comprise the disease prediction technique: pre-processing, feature extraction, and classification. In the pre-processing stage, excess and missing values are removed from the data set. The association between the target set and the characteristic will be established in the second phase. The entire set of data will be split into a training and test set during the final stage. For the prediction of chronic diseases, random forest, C4.5, and multilayer techniques are used. The ensemble classifier for the prediction of chronic illness will receive the output from each classifier. Using sensitivity, precision, recall, and F1 measure as metrics, the ensemble classification model's efficiency will be examined. The prediction of chronic diseases requires a reduction in the high complexity of the ensemble classifier.

This research work is based on the chronical disease prediction. Chronic disease prediction involves various steps including data input, extraction and classification. This research article describes the classification model. The proposed model is the combination of CNN and LSTM which is the main contribution of this paper. This is organized as section 1 give introduction and other information like need of study and research gaps. In the second section literature survey is proposed along with research methodology in section three. In the fourth section result and discussion is presented of the proposed model as well as existing model. In the last section conclusion and future directions for the chronical disease prediction is presented.

2. Literature Review

Predictive analytics in healthcare tries to forecast future events or outcomes pertaining to health based on clinical and/or nonclinical patterns in the data. Applications of data mining in pharmaceutical research have showed promise in tackling a variety of issues in drug discovery by reviewing the medical history of the patient, and delivers the best treatment for the patients by gaining knowledge from their symptoms and tests. In healthcare predictive analytics, study findings related to medical problems, hospital readmissions, therapy responses, and patient death are frequently of significant practical use. The creation of a healthcare predictive model entails two steps: first, the gathering of patient data in clinical trials using a predetermined set of methods. For instance, the Framingham Heart Study, the UK Prospective Diabetes Study, and the Lung Cancer Risk Prediction Model (UKPDS) [10]. Second, the use of already-collected patient information from clinical settings, such as EHRs, insurance claims, and clinical registries. A basic healthcare system's workflow for disease prediction consists of four successive layers: data collection, data fusion, feature extraction, pre-processing of the data, and disease prediction and recommendation.

H. A. Al-Jamimi, et.al (2024) proposed a new method that started with Feature Engineering and used a Support Vector Machine (SVM) in combination with RFE (Recursive Feature Elimination) [11]. To simplify the complexity of data, the introduced strategy detected and eliminated unnecessary features. The robust extreme Gradient Boosting (XGBoost) classifier known for its productiveness and skill in forecasting complicated relationships within the data was then fed to the processed dataset. The selected group learning algorithm described remarkable expertise in creating complicated design important for chronic disease prediction. By utilizing the Bayesian optimization, an important stage of optimization was proposed to increase the performance of model with the help of hyper parameter tuning. For the best predictive accuracy, this critically improving the model, steered the hyper parameter space and boosting the productivity of the search process. A significant improvement was depicted in early prediction of chronic diseases as depicted by the suggested model and illustrating the effectiveness of the proposed approach.

R. Ge, et. al (2020) developed an innovative ML-NN (multi-label neural network method) which was based on cross entropy loss function and backpropagation algorithm to forecast the chronic diseases combined with neural network and multi-level machine learning [12]. The introduced strategy attained constantly best in 5 performance measurements in comparison with 14 conventional multi-label study including 10 chronic diseases and 19,733 subjects. The suggested model constructively forecasted chronic diseases and helped the doctors to identify and treat the subjects as shown in the result.

S. M. M. Elkholy, et. al (2021) suggested an intelligent categorization and prediction system [13]. It employed altered DBNas (Deep Belief Network) classification algorithm to predict the diseases related to kidney, Softmax served as activation function and the Categorical Cross-entropy as loss function. In contrast to presenting model, this scheme can have forecasted the CKD with 98.5% accuracy and sensitivity of 87.5 % according to the calculation of the suggested model. A comparison can be made between the existing and proposed model in which proposed model obtained better performance with an accuracy of 98.52% and it also computed the productivity of suggested model. Hence, the proper predictor and classifier for CKD were presented by the suggested model. Result examination revealed that utilizing advanced deep learning methods were fruitful for the medical judgement and also assisted in prior prognosis of CKD and the connected stages that minimized the continuation of further harm in kidney.

G. Chen, et. al (2021) developed an AHDCNN (Adaptive Hybridized Deep Convolutional Neural Network) to detect the ability of several deep learning process for the initial identification of kidney disease skillfully and beneficially [14]. Effectiveness of the Classification technology relied on the role of the data set. With the help of CNN, an algorithm method had been developed to raise the validity of the classification system by decreasing the feature aspect. These high-level products help create organizational structures that differentiate between the two types of organizations. With the help of predictive examination, the IoMT (Internet of medical things platform) came to an end. In order to prove the predictive ability over the field of kidney disease, the machine learning advancements offered a guaranteed framework for the identification of intelligent solutions.

P. Chittora, et. al (2021) intended to forecast chronic kidney disease using the whole and significant elements of the CKD dataset [15]. Three different methods were used for feature selection: LASSO regression, Wrapper method and correlation-based feature selection. In this sense, seven classifier algorithms are used: Artificial Neural Network, C5.0, Logistic Regression, CHAID, LSVM (Linear Support Vector Machine), K-Nearest Neighbour and Random Tree. The outcome of each classifier were calculated using the following methods: all properties, CFS selected properties, Wrapper selected properties, LASSO-regression selected features, SMOTE with full properties and SMOTE with LASSO selective properties. It was seen that the highest accuracy of LSVM reached 98.86% in SMOTE with full properties.

R Halder, et. al (2024) proposed a machine learning-based kidney disease prediction (ML-CKDP) model with the dual objectives of enhancing dataset preprocessing for CKD classification and to create a web-based CKD prediction application [16]. The plan includes data augmentation before the process, conversion of categorical variables to numerical values, validation of missing data, and normalization via min-max scaling. Various methods such as correlation, chi-square, initial difference, repeated elimination, forward selection, lasso regression, and ridge regression were used for feature selection to validate the dataset. Seven classification methods, namely RF (Random Forest), AdaB (AdaBoost), GB (Gradient Boosting), XgB (XgBoost), NB (Naive Bayes), SVM (Support Vector Machine), and DT (Decision Tree) are used to predict CKD. The 100% accuracy achieved with RF (Random Forest) and AdaB (AdaBoost), is explained using various efficient methods such as 70:30, 80:20 data distribution and K fold set for 10 and 15.

J Zhang, et. al (2023) suggested that COPD (Chronic Obstructive Pulmonary Disease) was distinguished by determined respiratory disorders and airflow restriction [17]. AECOPD (Acute aggravation of COPD) was an acute worsening of respiratory symptoms, which required additional treatment and can result in intensify health status, enhancing the danger of hospitalization and death. Thus, early identification and diagnose complications of COPD was mandatory. This review discusses the current context of COPD exacerbations, current diagnostic tools and current biomarkers, and also includes the use of mobile medicine in the management of COPD to aid early detection and diagnosis.

R Sawhney, et. al (2023) presented the use of deep neural network-based multilayer perceptron classifiers has been proposed to treat CKD in patients [18]. The concept was studied using data from 400 individuals and various diseases and indicators were examined, including age, diabetes, red blood count, and more. Its main goals are to facilitate the introduction of deep learning to accurately diagnose CKD and learn from behavioral data. The main function of this study to treat chronic kidney disease is 100% accuracy of the deep neural network model. In addition to learning models such as support vector machines and naive Bayes classifiers. Neural models for the classification of chronic kidney disease may be a better choice for adaptive strategies because they can process complex data, calculate complicate data quantity collected from datasets, and use the arrangement of neuron within the network layers to adjust and learn about the essential information on their own.

Md. Ariful Islam, et. al (2023) discovered the relation presented among data attributes and the characteristics of the target class [19]. By using machine learning and predictive analytics, they were able to construct a set of prediction models. In addition to the class property, 25 variables were required in the beginning of the research The study also started with 25 different variables for certain features, but eventually narrowed the list down further, deciding that 30% of this limit was best for CKD diagnosis. Twelve different machine learning-based classification systems were tested in a supervised learning environment. In a field follow-up study, a total of 12 different machine learning-based classifications were examined, where the XgBoost classifier achieved 0.983 accuracy, 0.98 accuracy, 0.98 yield, and had an F1 score of 0.98, providing the best index. The approach of this study concludes that used predictive modelling in conjunction with recent advancements in machine learning presented an interesting way of discovering novel solutions, which subsequently used to examine the prediction accuracy in the area of kidney and beyond.

3. Research Methodology

The Chronical disease prediction model is designed which is based on the deep learning. Unlike present deep learning frameworks, the devised framework can forecast chronical disorders with a high degree of accuracy.

3.1. Phases of Proposed Model

The stages of the suggested approach are described as follows: -
3.1.1. Dataset input and pre-processing: - The dataset has been obtained from the UCI repository, and during the pre-processing stage, duplicate and missing values are eliminated.
3.1.2. Feature Reduction Phase: - The feature reducing process uses the PCA technique. The statistical technique called PCA (principal component analysis) transforms a set of linear variables that rely on the variables to form independent variables. PCA, also known as Orthogonal Linear Transformation, helps create an approximation of the original data set for different plots; the final goal is the projection of the first coordinate with at most the principal variable and the projection of the second coordinate with the second principal variable. difference. , Explains that the second part is perpendicular to the first part. To improve the variance of the data in the proposed domain, use PCA to perform a transformation expressed as $z = W_k^T x$, where $x \in R^d$ and $r < d$, in order to enhance the data variance within the projected space. The data matrix is specified as $X = \{x_1, x_2, \dots\dots.., x_i\}$, $x_i \in R^d$, $z \in R^r$ and $r < d$. The conversion is defined using a set of p-dimensional weight vectors $W = \{w_1, w_2, \dots\dots.., w_p\}$, where $w_p \in R^k$, which correspond to each x_i vector of X.

$$t_{k(i)} = W_{|(i)} T_{x_i}$$

To improve the difference , the initial weight $W1$ must meet the following conditions.

$$W_i = \arg max_{|w|} = \{\sum_i (x_i \cdot W)^2\}$$

A further expansion of the previous condition is provided as:

$$W_i = \arg max_{\|w\|=1} \{\|X. W\|^2\} = \arg max_{\|w\|=1} \{W^T X^T X W\}$$

. Since W is the corresponding eigenvector, a symmetric grid like the $X^T X$ can be efficiently analyzed after the largest eigenvalue of the matrix is attained. when W_1 is taken, the first element can be obtained by creating the first data matrix X for W_1 in the area provided by the transformation. The additional segments are obtained in this manner once the newly obtained components are subtracted.

3.1.3. Classification: - The proposed hybrid model is the combination of CNN and LSTM models. In this section work of CNN, LSTM and hybrid model is presented.

3.1.3.1. CNN Model: - CNNs are feedforward neural networks, meaning that there are no cycles or loops in the signal's transit through the network, which makes them appropriate for a range of image processing applications.

$$F(x) = f_N\left(f_{N-1}\left(\dots\left(f_1(x)\right)\right)\right)$$

The function in layer i is denoted by f_i in a normal CNN model, whereas N denotes the number of hidden layers. The key functional layers of a CNN model include the convolutional layer, activation layer, pooling layer, fully

connected layer, and prediction layer. Several convolution kernels $(g^1 \dots g^{k-1}, g^k)$ make up the convolutional layer's f. Each g^k in the k-th kernel denotes a linear function in the CNN model, and it can be expressed as [14]:

$$g^k(x, y) = \sum_{u=-m}^{m} \sum_{v=-n}^{n} \sum_{w=-d}^{w} W_k\,(u, v, w) I(x - u, y - v, z - w)$$

The positions of a pixel in the input image is denoted by (x, y, z). The weights for the k-th kernel are denoted by W_k, and the filter's height, breadth, and depth are represented by m, n, and w, respectively. f is a pixel-by-pixel non-linear function in the activation layer, like the rectified linear unit (ReLU) [11].

3.1.3.2. LSTM Model: - LSTM networks use micro-gate control to combine short-term and long-term memory and provide solutions for gradients that are missing within the bound. This network has three specific frameworks including an input gate, output gate and a forget gate. LSTM with the help of an architecture called gate in the cell state can remove and include information [9].

The primary step is concerned with deciding the information whose elimination is required from the cell state. For this purpose, a layer recognized as the forget gate is used. Following reading of the current input data (x_t) and the concealed state of the moment before it (h_{t-1}), the forget gate produces a vector between 0 and 1. The range of information that is rejected in the cell state c_t is indicated by the value of this vector, which lies between 0 and 1. A zero value indicates the rejection of all information, while 1 depicts the retainment of all information

$$z_f = \sigma\left(W_f.\,[h_{t-1}, x_t] + b_f\right)$$

The amount of new information to be added to the cell state is determined in the second step. This was achieved in two steps. First of all, the information to be updated is decided by means of an input gate operation usingh_{t-1} and x_t. Next, a novel candidate cell state z is obtained through a tanh layer using h_{t-1} and x_t.

$$z_i = \sigma(W_i.\,[h_{t-1}, x_t] + b_i)$$

$$z = \tanh(W.\,[h_{t-1}, x_t] + b)$$

Following expression is used to update the cell state:

$$c_t = z_f \odot c_{t-1} + z_i \odot z$$

Eventually, the output value is determined. The portion of the cell's state that provides output according to the inputs h_{t-1} and x_t is chosen after the cell state has been changed. [6]. It is required for inserting the input via a sigmoid layer termedas the output gate to get the decision conditions. After that, the cell state is passed via the tanh layer to get a vector within range [-1, 1]. The multiplication of decision conditions, that the output gate has provided when the output is forgotten, is done with it.

3.1.3.3. Hybrid Model: - This work suggests a hybrid deep learning framework that combines CNN and LSTM. A CNN is a large network that repeats and interprets information in a way that is comparable to the visual cortex of the brain. It has several hidden layers and is a deep neural network as well. Neural Networks are often used in the output layer of CNN for different types of classification. During the training phase, CNN uses a feature extractor rather than performing it by hand. CNNs that use feature extractors involve different types of neural networks, with the weights of each neural network being adjusted during training. The more layers used in the CNN's neural network to extract features, the better the image recognition, but at the expense of the previously difficult learning process, making CNNs less reliable and of poor quality. While another neural network classifies the features, a CNN neural network retrieves the features from the input images. The input image serves as the starting point for the feature extraction network. The extracted signals are used by the neural networks as

classification tool. This work uses a bi-directional paradigm for categorization. The Softmax Regression Layer couples the Forward and backward LSTMs of the Bidirectional LSTM (Bi-LSTM). The prediction effect is lost when dealing with time series since a regular LSTM network cannot learn all the sequences and lacks future context information.

3.2. Datasets of Chronical Disease

These three data are used to predict chronic diseases. The description of the document is an follows;-

1. Heart Disease Dataset: - The dataset for heart disease originates from https://archive.ics.uci.edu/ml/datasets/heart+disease and contains 76 attributes, but most published studies utilize only a subset of 14 attributes. The "objective" domain indicates the existence of heart disease in the patient, with values within the range [0,4].

2. Diabetic Dataset:- The second dataset was gathered from the following source: https://archive.ics.uci.edu/ml/datasets/diabetes. It is a dataset related to diabetes. Paper records and an automatic electronic recording system were the two sources from which the diabetes patient records were gathered. When it comes to time, automatic devices have internal clocks, but the data only records "hours" (e.g. lunch, dinner, and bedtime). Breakfast (8:00 AM), lunch (12:00 PM), dinner (6:00 PM), and bedtime (10:00 PM) are the times specified on the form. Therefore, electronic records have more accurate times than paper records, which can have inaccurate and inconsistent times.

3.Kidney Dataset:- The kidney dataset, which can be found at https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease, is the third dataset. Data containing 25 traits (including red and white blood cells) were collected in India over two months. The classification—which is either "ckd" or "notckd"—is the target; ckd stands for chronic renal disease. The file contains 400 row

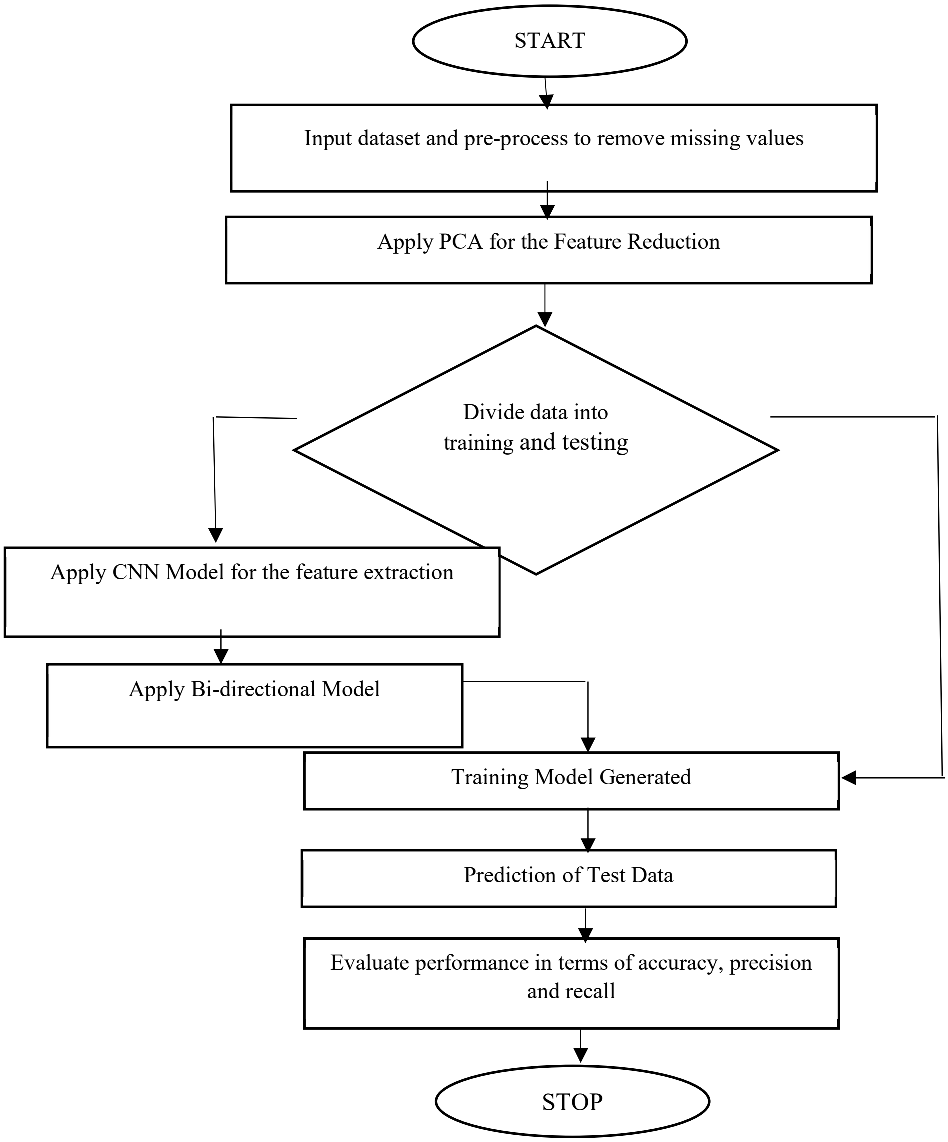

Figure1: Hybrid System Model

2. Precision: Precision also known as the positive predictive value is the percentage of relevant events in a simple that are identified or replicated in a standard, repeatable dataset or binary classification.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} - (2)$$

3. Recall: Recall measures the proportion of relevant events that are successfully recalled from each relevant event.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} - (3)$$

4.2. Diabetes Prediction Results

The prediction is based on the three chronic disease datasets. Diabetes is among the conditions that affect the heart. The suggested framework is used to predict diabetes and contrasted with other categorization methods. These methods include Logistic regression, K-nearest neighbor, Support vector machine, Naïve Bayes, decision tree and Random forest in terms of accuracy, precision and recall.

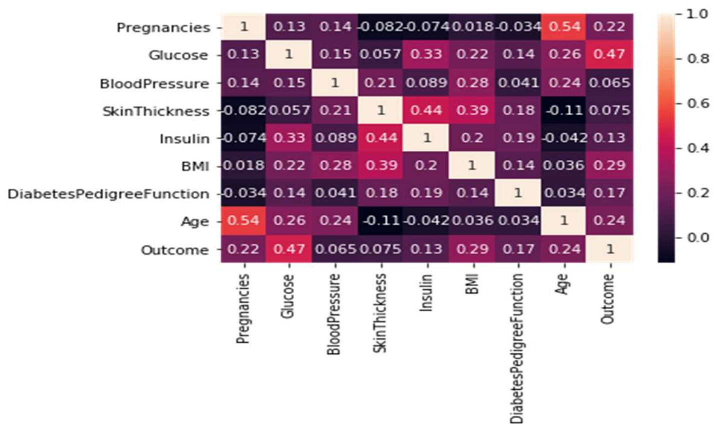

Figure 2: Correlation Matrix

The different dataset attributes are presented in Figure 2, along with the relationship between each attribute and the other attributes. The correlation matrix is plotted which define relationship between attributes of the dataset.

Table 1: Analysis of Performance Using the Diabetes Dataset

Model	Accuracy	Precision	Recall
Logistic Regression	71.42	73	72.2
KNN	78.37	79	78
SVM	73.37	75	74
Naïve Bayes	71.42	72	71
Decision Tree	68.18	69	68
Random Forest	75.17	76	75
Proposed Model	83.76	85	83

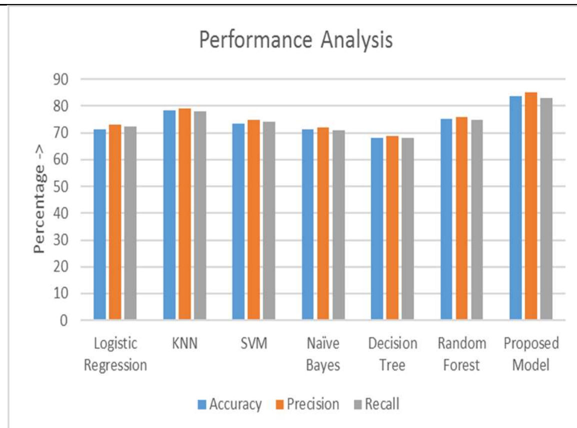

Figure 3: Analysis of Performance Using the Diabetes Dataset

Figure 3 illustrates how the suggested model's efficacy for the diabetic prediction is compared to that of other classification models. With a precision of 73%, recall of 72.2%, and accuracy of 71.42%, Logistic Regression was employed. Better results were obtained by the K-Nearest Neighbors (KNN) model, which had a 78.37% accuracy, 79% precision, and 78% recall. The accuracy, precision, and recall of the Support Vector Machine (SVM) were 73.37%, 75%, and 74%, respectively. Naïve Bayes had comparable performance to Logistic Regression, with 71.42% accuracy, 72% precision, and 71% recall. With an accuracy of 68.18%, precision of 69%, and recall of 68%, the Decision Tree model lags behind. With an accuracy of 75.17 percent, precision of 76%, and recall of 75%, the Random Forest model performed better. With an astounding 83.76% accuracy, 85% precision, and 83% recall, the suggested model fared better than any of the others.

3.2. Heart Disease Prediction

Heart related disorders ranks second among the major chronic illnesses. The proposed model is tested on the heart disease prediction. The performance of proposed model is also compared with other models in terms of accuracy, precision and recall

Table 2: Analyzing Performance to Predict Heart disorders

Model	Accuracy	Precision	Recall
Logistic Regression	85.25	86	85
KNN	67.21	67	67
SVM	81.97	83	81
Naïve Bayes	85.25	85	84
Decision Tree	81.97	82	81
Random Forest	95.08	96	95
Proposed Model	98.37	98	98

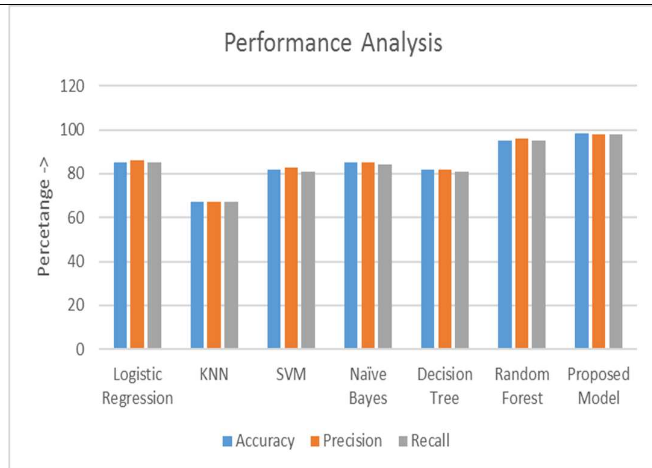

Figure 4: Analyzing Performance to Predict Heart disorders

The suggested model's effectiveness is evaluated using the heart disease dataset, as seen in figure 4. With a precision of 86%, recall of 85%, and accuracy of 85.25%, Logistic Regression was employed. The K-Nearest Neighbors (KNN) model performed worse, with 67.21% accuracy, 67% precision, and 67% recall. Strong performance was shown by the Support Vector Machine (SVM), which had 81.97%, accuracy, 83% precision, and 81% recall. Naïve Bayes also performed well, matching Logistic Regression with an accuracy of 85.25%, precision of 85%, and recall of 84%. The Decision Tree model achieved an accuracy of 81.97%, precision of 82%, and recall of 81%. The Random Forest model excelled with an accuracy of 95.08%, precision of 96%, and recall of 95%. The framework suggested in thus work surpassed all other models significantly, with an outstanding 98.37 accuracy, precision of 98%, and recall of 98%.

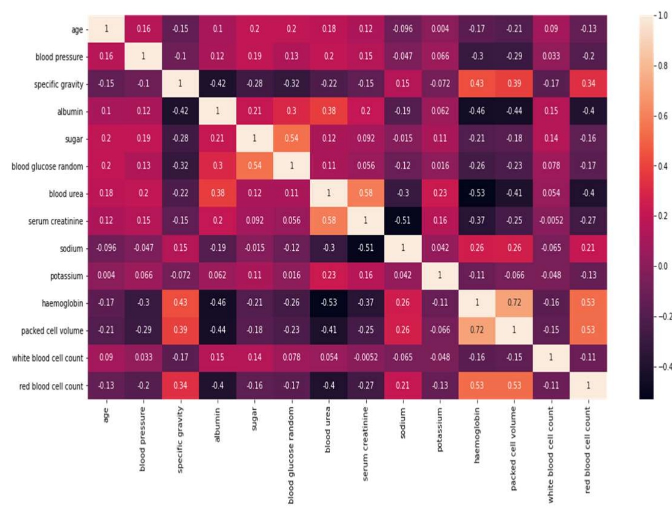

Figure 5: Attributes of Dataset

Figure 5 illustrates the different features associated with renal illness as a confusion matrix. The matrix displays how different traits relate to one another.

3.3. Analyzing Performance to Predict Kidney Disease

Model	Accuracy	Precision	Recall
Logistic Regression	82	82	83
KNN	65.09	64	65
SVM	78.18	78	79
Naïve Bayes	82.16	82	81
Decision Tree	80	82	81
Random Forest	93.02	93	92
Proposed Model	95	90.25	97.34

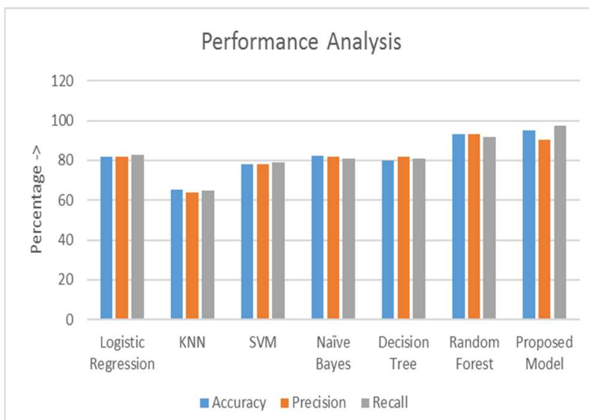

Figure 6: Analyzing Performance to Predict Kidney Disease

The effectiveness of the suggested approach is evaluated for the prediction of renal illness, as seen in figure 6. With 82% precision, 82% recall, and 83% accuracy, Logistic Regression performed well. KNN yielded poorer results, with recall of 65%, accuracy of 65.09%, and precision of 64%. SVM demonstrated strong performance, with 78.18% accuracy, 78% precision, and 79% recall. With an accuracy of 82.16%, recall of 81%, and precision of 82%, Naïve Bayes performed well. DT showed 80% accuracy, 82% precision, and 81% recall. With 93% precision, 92% recall, and 93.02% accuracy, the Random Forest model performed exceptionally well. The suggested model demonstrated superior performance compared to the other models, with remarkable 95% accuracy, 90.25% precision, and 97.34% recall.

Conclusion

The main issue with this sickness in the modern era is that patients with chronic illnesses don't exhibit any symptoms. These individuals spread the illness to others worldwide and are the carriers of the sickness. Chronic diseases can arise for a variety of reasons. The first explanation is the early stage of the pandemic breakout and the lack of information openness. Pre-processing, feature extraction, and classification are some of the phases involved in the chronic illness prognosis process. A hybrid approach is suggested for the prognosis of chronic illnesses. CNN and bidirectional LSTM are combined to create the suggested model. Bidirectional LSTM is utilized for classification, and the CNN model is employed for feature extraction. The preparation is verified in three different diseases (diabetes, kidney disease and heart disease). It is analysed from the results that proposed model achieve 83 percent accuracy on diabetic dataset 98 percent accuracy on heart disease dataset and 95 percent on kidney dataset. It is analysed concluded that proposed model achieves 10 percent high accuracy on all three datasets than existing machine learning models. In future proposed model can be further extended with class balancing algorithms.

References

[1] Q. Zhao, J. Li, L. Zhao and Z. Zhu, "Knowledge Guided Feature Aggregation for the Prediction of Chronic Obstructive Pulmonary Disease with Chinese EMRs," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 6, pp. 3343-3352, Nov.-Dec. 2023, doi: 10.1109/TCBB.2022.3198798.

[2] Shahid Mohammad Ganie, Pijush Kanti Dutta Pramanik, "A comparative analysis of boosting algorithms for chronic liver disease prediction", Healthcare Analytics, 2024, Vol. 5, 100313, doi.org/10.1016/j.health.2024. 100313..

[3] R Mohebi, Roland van Kimmenade, Cian P. McCarthy, Craig A. Magaret, Grady Barnes, Rhonda F. Rhyne, Hanna K. Gaggin, James L. Januzzi, "Performance of a multi-biomarker panel for prediction of cardiovascular event in patients with chronic kidney disease", International Journal of Cardiology, 2023, Vol. 371, Pages 402-405, doi.org/10.1016/j.ijcard.2022.09.074.

[4] D Jain, Vijendra Singh, "Feature selection and classification systems for chronic disease prediction: A review", Egyptian Informatics Journal, 2018, Vol.19, Issue 3, Pages 179-189, doi.org/10.1016/j.eij.2018.03.002.

[5] Amir Rehman, Huanlai Xing, Mehboob Hussain, Nighat Gulzar, Muhammad Adnan Khan, Abid Hussain, Sajid Mahmood, "HCDP-DELM: Heterogeneous chronic disease prediction with temporal perspective enabled deep extreme learning machine", Knowledge-Based Systems, 2024, Vol. 284,111316, doi.org/10.1016/j.knosys.2023.111316.

[6] Heidi T. May, Jeffrey L. Anderson, Joseph B. Muhlestein, Kirk U. Knowlton, Benjamin D. Horne, "Intermountain chronic disease risk score (ICHRON) validation for prediction of incident chronic disease diagnoses in an australian primary prevention population", European Journal of Internal Medicine, 2020, Vol. 79, Pages 81-87, doi.org/10.1016/j.ejim.2020.06.009.

[7] H. H. Yördan, M. Karakoç, E. Çalğici, D. Kandaz and M. K. UÇar, "Hybrid AI-Based Chronic Kidney Disease Risk Prediction," *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Sivas, Turkiye, 2023, pp. 1-4, doi: 10.1109/ASYU58738.2023.10296642.

[8] O. Aruna and S. Sameerunnisa., "Chronic Kidney Disease Prediction using Data Pre-Processing Techniques," *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2023, pp. 1665-1670, doi: 10.1109/ICSSIT55814.2023.10061025.

[9] C. P. Kashyap, G. S. Dayakar Reddy and M. Balamurugan, "Prediction of Chronic Disease in Kidneys Using Machine Learning Classifiers," *2022 1st International Conference on Computational Science and Technology (ICCST)*, CHENNAI, India, 2022, pp. 562-567, doi: 10.1109/ICCST55948.2022.10040329.

[10] K Chien, Hung-Ju Lin, Bai-Chin Lee, Hsiu-Ching Hsu, Yuan-Teh Lee, Ming-Fong Chen, "A Prediction Model for the Risk of Incident Chronic Kidney Disease", The American Journal of Medicine, 2010, Vol. 123, Issue 9, 2010, Pages 836-846.e2, doi.org/10.1016/j.amjmed.2010.05.010.

[11] H. A. Al-Jamimi, "Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction," in *IEEE Access*, vol. 12, pp. 62215-62233, 2024, doi: 10.1109/ACCESS.2024.3395512.

[12] R. Ge, R. Zhang and P. Wang, "Prediction of Chronic Diseases with Multi-Label Neural Network," in *IEEE Access*, vol. 8, pp. 138210-138216, 2020, doi: 10.1109/ACCESS.2020.3011374.

[13] S. M. M. Elkholy, A. Rezk and A. A. E. F. Saleh, "Early Prediction of Chronic Kidney Disease Using Deep Belief Network," in *IEEE Access*, vol. 9, pp. 135542-135549, 2021, doi: 10.1109/ACCESS.2021.3114306.

[14] G. Chen *et al*., "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," in *IEEE Access*, vol. 8, pp. 100497-100508, 2020, doi: 10.1109/ACCESS.2020.2995310.

[15] P. Chittora *et al*., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in *IEEE Access*, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[16] R Halder, Mohammed Nasir Uddin, Md. Ashraf Uddin, Sunil Aryal, Sajeeb Saha, Rakib Hossen, Sabbir Ahmed, Mohammad Abu Tareq Rony, Mosammat Farida Akter, “ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application”, Journal of Pathology Informatics, 2024, Vol. 15, 100371, doi.org/10.1016/j.jpi.2024.100371.

[17] J Zhang, Fangman Chen, Yongli Wang, Yahong Chen, “Early detection and prediction of acute exacerbation of chronic obstructive pulmonary disease”, Chinese Medical Journal Pulmonary and Critical Care Medicine, 2023, Vol. 1, Issue 2, Pages 102-107, doi.org/10.1016/j.pccm.2023.04.004.

[18] R Sawhney, Aabha Malik, Shilpi Sharma, Vipul Narayan, “A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease”, Decision Analytics Journal, 2023, Vol. 6, 100169, doi.org/10.1016/j.dajour.2023.100169.

[19] Md. Ariful Islam, Md. Ziaul Hasan Majumder, Md. Alomgeer Hussein, “Chronic kidney disease prediction based on machine learning algorithms”, Journal of Pathology Informatics, 2023, Vol. 14, 100189, doi.org/10.1016/j.jpi.2023.100189.