

Chronic kidney disease prediction model using machine learning approach

Munusamy Chitra, Abdul Kuthus Parveen, Murugadoss Elavarasi, Jayamoorthy Sangeetha, Ramalingam Vaithilingame

Department of Computer Applications, Perunthalaivar Kamarajar Arts College, Puducherry, India

Article Info

Article history:

Received Mar 29, 2021

Revised Oct 5, 2021

Accepted Oct 27, 2021

Keywords:

Chronic kidney disease

Classification methods

Health care

Machine learning

Medical code data

ABSTRACT

Chronic disease (CD) such as kidney disease and causes severe challenging issues to the people all around the world. Chronic kidney disease (CKD) and diabetes mellitus (DM) are considered in this paper. Predicting the diseases in earlier stage, gives better preventive measures to the people. Healthcare domain leads to tremendous cost savings and improved health status of the society. The main objective of this paper is to develop an algorithm to predict CKD occurrence using machine learning (ML) technique. The commonly used classification algorithms namely logistic regression (LR), random forest (RF), conditional random forest (CRF), and recurrent neural networks (RNN) are considered to predict the disease at an earlier stage. The proposed algorithm in this paper uses medical code data to predict disease at an earlier stage.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Munusamy Chitra

Department of Computer Applications, Perunthalaivar Kamarajar Arts College

Kalitheerthalkuppam, Puducherry 605107, India

Email: chitra.jrf@gmail.com

1. INTRODUCTION

Healthcare system is the maintenance or improvement of health of a person in right manner. Health of a person depends on two categories such as mental and physical health. Mental health keeps the person happy and the physical health makes the person to do various physical activities without any difficulties. The preventive measures are early diagnosis of the disease, regular treatment to the patient and cure and recovery from the disease. Health care provider plays major role in health care domain (HCD). The main focus of healthcare provider is to concentrate more on preventive measures rather than treating the patients [1], [2].

Healthcare issue can be prevented by early diagnosis technique. Diagnosing disease at the initial stage provides better improvement and can able to identify the severity of the disease. Due to cost effective factor of disease prediction, people from the poor financial background not able to do at any instant of time. Hence a simple, reliable disease prediction algorithm is needed to identify the risk factors for future health improvement. Machine learning (ML) algorithms in HCD entirely based on medical data collected at real time or secondary data collected from various sources to predict accurate outcome [3].

Several chronic disease (CD) namely heart disease, cancer, and respiratory problems exist for a long period of time throughout a year or several years under follow up by the healthcare providers. CD causes major health issues to the individual which is the major reason for high death rate in all over the world. The "centre for disease control and prevention" states CD can be prevented at the start stage before going to the severe condition. An early detection or prediction of CD helps to reduce the severity of the disease also cost effective to an individual [4], [5]. This paper concentrates on two CD's namely chronic kidney disease (CKD) and diabetes mellitus (DM). DM is also called as type II diabetes.

The CKD is a major issue which fails the function of kidneys due to unwanted blood wastage storage in kidneys. At the earlier stage these diseases do not show any symptoms but at the later stage causes severe problem to the individual. The symptoms are tiredness, vomiting, and upset stomach. The disease at the final stage leads to kidney failure which needs kidney transformation for proper function of the human body. In order to avoid these major issues preventive measures are needed for the welfare of the society. Hence, this paper discusses about disease prediction algorithm to reduce high risk rate of the individual [6].

The second type of disease considered in this paper is DM also called as type II diabetes. Type II is another type of CD in which blood glucose level exceeds at higher rate due to defective insulin production or action of human body. Similar to CKD, DM also leads several health issues such as kidney disease, heart disease and blindness. In order to overcome these issues, health care modeling is proposed in this paper. Proposed ML model for both CKD and DM helps to identify undiagnosed individuals and directs allow them to follow some preventive measures such as to control blood pressure (BP) and glucose level [7], [8].

- Medical data

Predicting the disease at the earlier stage is a challenging issue; the case in which the diseases without showing any symptoms to the patients. Various technological measures are required to predict the severity of the disease. Based upon the measures taken from the patient, health care providers can give proper treatment at the earlier stage. HCD is most important for the welfare of people. Healthcare dataset plays major role in HCD to estimate disease level at the initial stage. Healthcare dataset provides more useful and intelligent information to the people [9] to identify the disease at the initial stage. Some of the various categories of medical datasets used in the literature are shown in Table 1 found in electronic health records (EHRs) and insurance claims records of the patients.

Table 1. Different types of medical data

S. No.	Medical data	Characteristics of medical data	Examples	Methods used	Technology used
1	Simple variables	Numerical and categorical	Age, sex, and ethnicity height, weight, BP, and pulse	Analytical and statistical methods	X-ray, ultrasound, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI)
2	Image data	Radiology	Images of internal tissues	visual diagnostics by an expert	Computer vision and natural language processing (NLP) to recognize text and extract the meaning of the text
3	Freetextdata	Computer vision	Incomplete sentences, or detailed paragraphs	Computer vision	NLP
4	Coded data	Computational processing	Coded data	Coding system	human-defined categories
5	Diagnosis codes	A patient visit doctor directly to receive a primary diagnosis code to indicate the presence of disease	Secondary data	From the existing records	The primary diagnosis code indicates the primary reason to visit the doctor and secondary codes record also part of doctor's note
6	Procedure codes	Surgery, physiotherapy, and diagnostic interventions	Alphanumeric strings of five length	Surgery, physiotherapy, and diagnostic interventions	Healthcare common procedure system
7	Drug codes	Record the brand name of the drug	Brand name of the drug	National drug code set	United States adopted name system

- Machine learning in healthcare

The ML is a method of developing a model based on mathematical, computational, and statistical method to find patterns and to extract information from the existing data. ML research transforms human life in systematic way with the use of intelligent technology. In HCD, ML models aims to provide artificial support and help to health care providers. It objective is to replace human doctors, nurses and others where humans found difficulties and struggle [10] for man power, infrastructure support and technical issues. The performance of the ML models is improved by analyzing various factors such as proper healthcare datasets for example an incorrect training data yields incorrect result. Hence, data set is the heart of ML techniques. A high quality training data set is most important to get accurate result [11]. The following are some of the characteristics of healthcare dataset: i) imbalanced data set: certain important properties of the dataset have very small proportion of the dataset; ii) sample size: data set size is most important to get the accurate result; iii) realism: it defines an originality of the dataset used; and iv) generality: healthcare datasets need to have most important features related to a particular domain.

2. RELATED WORK

The importance of ML models for disease prediction in healthcare domain is discussed briefly in this section. The two popular predictive techniques namely logistic regression (LR) and the Cox proportional hazards model are commonly used by most of the authors in this research field. Echouffo-Tcheugui and Kengne [11] reviewed 30 CKD prediction models from the existing research work since 1980s. Further examples are found in the papers [12].

Research by Cruz and Wishart [13] discusses about a few ML techniques to predict cancer prognosis using decision trees, neural networks, and nearest neighbour classifiers and in paper. Research by Conroy *et al.* [14] build a Weibull hazards model with clinical and demographic data to predict risk rate of cardiovascular disease. Research by Khalilia *et al.* [15] uses a random forest (RF) algorithm using highly imbalanced data.

Topic modelling techniques based on latent dirichlet allocation (LDA) and its variants are most popular in the recent years which use both freetext and medical code data. According to Halpern *et al.* [16] investigates both supervised and unsupervised ML models for dimensionality reduction (DR) and feature extraction (FE) methods for clinical data prediction purpose. In most cases, these models are applied in emergency department as freetextdata and use the learned topic distributions to predict patient risk rate to develop sepsis and allow the patient to admit in intensive care unit (ICU).

According to Lehman *et al.* [17] rain a topic model on unified medical language system (UMLS) codes which is extracted from unstructured nurse notes to predict hospital mortality. Their prediction shows that learned topic weights with a patient improves traditional risk factors of the disease. Research by Dritsas and Trigka [18] use a topic model on freetext from EHRs to explore infant colic issues in depth. They hypothesize that the learned topics automatically assigns cases of infant colic from EHRs in the absence of the word “colic” in the record. Research by Luo *et al.* [19] applies topic modelling to international classification of diseases (ICD)-10 codes to summarize clinical information and generates medical research hypotheses. Research by Perotte *et al.* [20] uses supervised learning model with hierarchical labels to automatically assign ICD-9 codes to patient discharge summary list.

Research by Wang *et al.* [12] build an unsupervised model based on Markov chain model. This model identifies the disease level and its associated complaints at each stage. Research by Sun *et al.* [21] evelops a model using the factors knowledge, data and ICD-9 codes for heart failure prediction. Research by Singh *et al.* [22] uses hierarchical structure of ICD-9 codes to develop feature vector representations of patients. Research by Tsui *et al.* [23] use ICD-9 codes to develop an epidemic detection algorithm. Research by Davis *et al.* [24] develops a model using collaborative filtering approach.

3. PROPOSED DISEASE PREDICTION MODEL

The proposed ML model in this paper aims to predict CD using different datasets. Dataset plays major role ML models without data these models will not function. This paper considers two healthcare datasets from different regions of United State (US) with various features. These two datasets are named as D1 and D2, respectively. Along with the basic features, these datasets also contain the special feature namely insurance claims records of people for several years. This record indicates that the insurance enrolment status of each person throughout the year which helps to determine whether the patient’s insurance policy coverage is continuous or not. It is important to observe coverage of policy by an individual diagnosed with some target disease during a particular time or not. The size of D1 is considerably larger than size of D2.

In order to analyze the dataset, consider the following assumptions: let N_o be an observation period and N_f be the follow-up period in years. The various types of medical code data used are diagnosis codes, procedure codes, drug codes. For each person in the dataset construct a prepared dataset with two properties with the following constraints:

- Constraint 1: let us assume that if the person has a continuous insurance enrollment period for the observation and followup period from N_o+N_f consecutive years.
 - a. File the codes which occurs in the first observation period (N_o) years as the first continuous enrollment period.
 - b. When the person diagnosed the target disease before the end of the observation period leaves that person from the file.
 - c. When the person diagnosed with the target disease within the follow-up period (N_f) years at the end of the follow-up time then set the target variable as 1.
 - d. When the person not diagnosed with any the target disease at any point of time before the follow-up period set the target variable as 0.
- Constraint 2: fails of constraint 1
 - a. Leave this person out.

This research work is carried out to predict target diseases such as CKD and DM. ICD-9 diagnosis codes are used to determine disease diagnosis as shown in Table 2. These codes address the following two

questions: i) whether the given medical codes for an undiagnosed individual collected during an observation period of N_o years? and ii) can we predict the patient diagnosed with the target disease during the follow-up period of N_f years? In order to answer the above-mentioned questions, the following section describes the training, testing and validation process of the proposed ML model.

Table 2. ICD-9 codes for CKD and DM

Target disease	ICD-9 codes
CKD	403.x, 404.x, 582.x, 583.x, 585.x, 586.x, 588.0
DM	250.x0, 250.x2

3.1. Training process

The main functionality of ML models is to use dataset to take decisions or to predict the final outcome of the proposed model. ML algorithms construct models from an observed data through training process. Training process consists of selection of a model which best “fits” the training data set. In classification process the training data set is a collection labels which is a subset of $X \times Y$ in supervised learning model.

3.2. Testing process

In training process, construct a classifier which fits training set of labelled datasets under supervised learning. The main aim of good classifier is to perform well in all appropriate dataset. The good classifier performs well in new and unseen data. The performance of the proposed model is measured in terms of accuracy, receiver operating characteristic (ROC) and area under curve (AUC) curves as follows:

3.2.1. Accuracy

Accuracy is one of the metrics for hard classifier (f). It is defined as the percentage of correct predictions in the test dataset to total number of predictions. This includes true positive, true negative, false positive and false negative.

$$\text{Accuracy} = \text{Number of Correct Predictions} / \text{Total Number of Predictions} \quad (1)$$

3.2.2. Receiver operating characteristic and area under curve

The two popular performance metrics for soft classifier are ROC curve and the AUC. The ROC curve shows the performance by visual representation whereas AUC shows the measures in numerical representation. The measurement values of both ROC curve and AUC values are shown in Figure 1. True positive rate (TPR) is defined as (2):

$$\text{TPR} = \text{True Positive} / (\text{True Positive} + \text{False Negative}) \quad (2)$$

False positive rate (FPR) is defined as (3):

$$\text{FPR} = \text{False Positive} / (\text{False Positive} + \text{True Negative}) \quad (3)$$

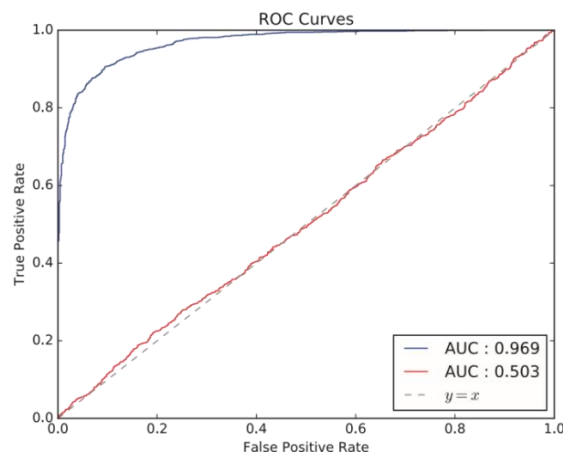


Figure 1. ROC curves and AUC scores for a good classifier (blue) and a poor classifier (red)

The ROC is a type of parameterized curve and its test set is finite. ROC curves plotted linearly between neighboring points and consists of unit square $[0,1] \times [0,1]$. A positive slope near the origin indicates good classifier and a gentler slope lies close to the line $y=x$ indicates poor classifier. The AUC portion is bounded by the bottom of the unit square and the ROC curve. The AUC score near to '1' and '0.50' indicates good and poor performance respectively.

4. RESULTS AND EXPERIMENTAL ANALYSIS

This section shows that the results and analysis of the disease prediction model in detail. This research work carried out three different types of experiments. The trials were such as: i) finding the best classifier; ii) inter-population prediction; and iii) predicting over variable follow-up periods.

4.1. Finding the best classifier

This experiment uses D1 and ICD-9 diagnosis codes to train and analyze the performance of four classifiers namely LR models, RF, conditional random forest (CRF) and recurrent neural networks (RNN) with various data representations. From the four classifiers, best classifier is predicted for both CKD and DM with the observation period $N_o=\{1,2\}$ and the follow-up period $N_f=2$. For each target disease namely CKD and DM, randomly divide the dataset into three categories such as 50% of the data is used for training phase, 25% is used for validation phase and remaining 25% is used for testing phase. This is basic manner to partition the dataset to find out diseased to non-diseased cases.

The validation dataset is used for hyper parameter optimization (HPO) with the set conditions to optimize each classifier. The type and weight (TW) of the regularization in LR, the number of trees (NT) in a RF, the window size (WS) of feature functions in a CRF and the size of the layers (SL) in an RNN. These conditions are named as hyper parameters (HP). The following four classifiers and data representations are considered in this paper.

The LR models: LR is trained using three data representations: i) the simple bag of word (BOW) representation (LR+BOW); ii) the lower-dimensional representation given by the carbon capture and storage (CCS) mapping (LR+CCS); and iii) the LDA topic distribution representation (LR+LDA). For these three cases, include ' L_2 ' aspenalty and optimize the penalty weight ' C ' for the set $\{0.1, 0.5, 1.0, 5.0\}$. For LR+LDA, optimize the number of topics in the LDA model for the given set $\{10, 30, 50, 100\}$.

The RF models: similar to LR model, the same data representation and pair of notations are used to train RF classifier such as RF+BOW, RF+CCS, and RF+LDA. For each case, optimize NT in the forest for the set $\{50, 100, 250, 500\}$. For RF+LDA, optimize the number of topics over the set $\{10, 30, 50, 100\}$.

The CRF models: this model is trained using window features which consist of ICD-9 codes and the CCS groupings and bias term and WS is optimized for the set $\{0, 1, 3\}$. The RNN models: RNN is trained for both ICD-9 sequence code and CCS symbolic sequence (RNN+CCS). RNN consists of four layers such as an input, a gated recurrent unit (GRU) recurrent, a dense feedforward and a double dimensional softmax output layer. RNN is trained with two training epochs finally optimize the number of neurons in both the recurrent layer and the feedforward layer over the set $\{128, 256\}$. The AUC scores of given models for the disease prediction using only ICD-9 diagnosis code data is shown in Table 3.

Table 3. ICD-9 codes with CKD and DM

Disease	Method	AUC ($N_o=1, N_f=2$)	AUC ($N_o=2, N_f=2$)
CKD	LR+LDA	0.776	0.801
	RF+CCS	0.756	0.813
	CRF	0.590	0.617
	RNN+CCS	0.812	0.842
DM	LR+LDA	0.757	0.766
	RF+CCS	0.714	0.753
	CRF	0.649	0.582
	RNN+CCS	0.792	0.803

4.1.1. Discussion

Among the several classifiers considered in this paper, the performance of RNN with CCS representation of the ICD-9 code sequences is better in all cases for the target diseases CKD and DM during the two year long observation periods. RNN approach outperforms CKD prediction than DM prediction. RNN classifier consists of feed forward layers of size 256. Also, this can be further improved by increasing the number of neurons and the optimal number of neurons in the recurrent layer ranges from 128 to 256. Table 4 shows the result of RNN classifier which includes results of hyper-parameters of good classifiers. Figure 2 shows TPR and FPR for CKD and DM.

Table 4. Hyper-parameters for the best classifiers

Disease	Hyper-parameters ($N_o=1$)	Hyper-parameters ($N_o=2$)
CKD	Method: RNN+CCS	Method: RNN+CCS
	Feedforward layer size: 256	Feedforward layer size: 256
	Recurrent layer size: 256	Recurrent layer size: 128
DM	Method: RNN+CCS	Method: RNN+CCS
	Feedforward layer size: 256	Feedforward layer size: 256
	Recurrent layer size: 128	Recurrent layer size: 256

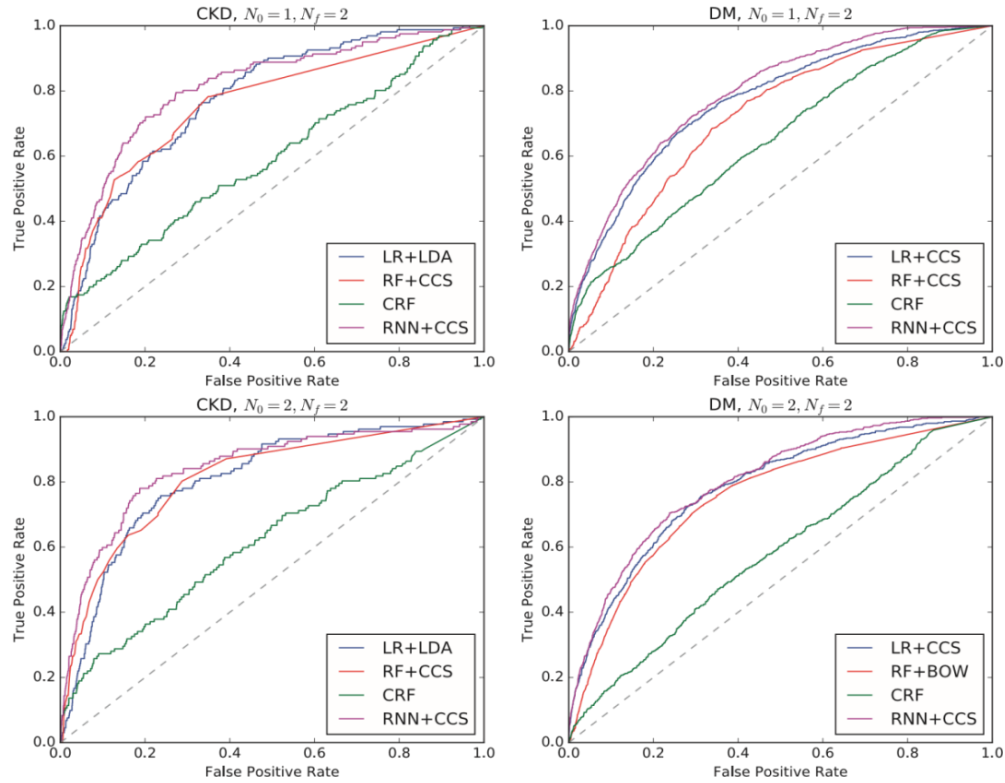


Figure 2. TPR and FPR

4.2. Inter-population prediction

This experiment trains a classifier on D1 and tested D2, and vice versa. Also, trained using ICD-9 diagnosis codes and both datasets using $N_o=1$ and $N_f=2$. Similar to experiment 1, RNN+CCS method with recurrent and feed forward layers both of size 256 used and trained using two epochs. Figure 3 shows AUC curve for target diseases CKD and DM.

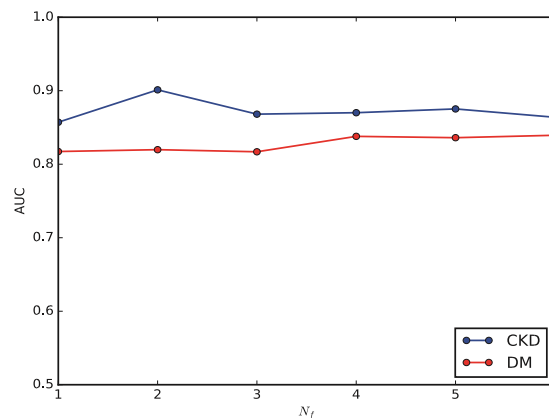


Figure 3. AUC curve

4.2.1. Discussion

The outcome of an inter-population prediction experiment for both CKD and DM with RNN+CCS classifier is shown in Table 5. AUC scores for RNN+CCS classifier in an inter-population prediction experiment for both CKD and DM ranges from 0.5 to 1. AUC scores of CKD and DM for several followup times have the best AUC scores as shown in Figure 3. The outcome of this experiment is lower than the original population. This is due to different geographical location of the people. People living in rural and urban areas will have different life style and hygienic condition. Also, medical coding practices may change from one hospital to another hospital. Hence the result shows that variance in the outcome of the given dataset.

Table 5. Inter-population prediction

CKD	DM
0.765	0.653
0.778	0.732

4.3. Predicting over variable follow-up periods

This experiment investigates the changes in the follow-up period ' N_f ' from $N_f=1$ to $N_f=6$. RNN+CCS use 256 neurons in the feedforward layer and 128 neurons in the recurrent layer. The dataset is randomly partitioned into training dataset (75%) and test dataset (25%) with two epochs. Table 6 shows AUC scores each followup time N_f .

Table 6. Result of variable follow-up periods

Disease	$N_f=1$	$N_f=2$	$N_f=3$	$N_f=4$	$N_f=5$	$N_f=6$
CKD	0.842	0.901	0.868	0.870	0.875	0.864
DM	0.803	0.820	0.817	0.838	0.836	0.839

4.3.1. Discussion

The AUC score of CKD using RNN+CCS classifier for $N_o=N_f=2$ increases from 0.842 to 0.901. For DM, the AUC score increases from 0.803 to 0.820. AUC score indicates that it can predict disease for long term efficiently in this case follow-up period ranges from 1 to 6 years.

5. CONCLUSION

This research work was successfully carried out to analyze the performance of CKD prediction with medical code data of an individual (patient). The objective of this is to use ICD codes to predict target disease such as CKD and DM. ICD-9 codes are used for feature selection or DR. Four different classification algorithms are used in this paper namely LR, RF, CRF, and RNN. Compared to other classifiers the performance of RNN provides better solution. Hence this paper analyzed the performance of CKD and DM from the medical coded dataset with different ML classifiers for different follow-up period. In future, needs to add more additional data to the classifier to find the target disease.




REFERENCES

- [1] M. A. Abdel-Fattah, N. A. Othman, and N. Goher, "Predicting chronic kidney disease using hybrid machine learning based on apache spark," *Computational Intelligence and Neuroscience*, pp. 1–12, Feb. 2022, doi: 10.1155/2022/9898831.
- [2] S. Pal, "Chronic kidney disease prediction using machine learning techniques," *Biomedical Materials & Devices*, pp. 185–190, Aug. 2022, doi: 10.1007/s44174-022-00027-y.
- [3] R. S. S. Shaji, S. R. Ajina, V. P. S. R., and J. A., "Chronic kidney disease prediction using machine learning," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 7, pp. 137–140, Jul. 2020, doi: 10.17577/IJERTV9IS070092.
- [4] R. A. S. P. S., K. V. Rangarao, and A. Saranya, "Efficient datamining model for prediction of chronic kidney disease using wrapper methods," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 8, no. 2, pp. 63–70, Aug. 2019, doi: 10.11591/ijict.v8i2.pp63-70.
- [5] T. K. Chen, D. H. Knicely, and M. E. Grams, "Chronic kidney disease diagnosis and management: a review," *JAMA*, vol. 322, no. 13, pp. 1294–1304, Oct. 2019, doi: 10.1001/jama.2019.14745.
- [6] R. G. Nelson *et al.*, "Development of risk prediction equations for incident chronic kidney disease," *JAMA*, vol. 322, no. 21, pp. 2104–2114, Dec. 2019, doi: 10.1001/jama.2019.17379.
- [7] Z. Wang, J. W. Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, "Machine learning-based prediction system for chronic kidney disease using associative classification technique," *International Journal of Engineering & Technology*, vol. 7, no. 4.36, pp. 1161–1167, Dec. 2018, doi: 10.14419/ijet.v7i4.36.25377.




- [8] J. O. Hero, R. J. Blendon, A. M. Zaslavsky, and A. L. Campbell, "Understanding what makes Americans dissatisfied with their health care system: an international comparison," *Health Affairs*, vol. 35, no. 3, pp. 502–509, Mar. 2016, doi: 10.1377/hlthaff.2015.0978.
- [9] H. Schmidt *et al.*, "Chronic disease prevention and health promotion," in *Public Health Ethics: Cases Spanning the Globe*, Cham: Springer International Publishing, 2016, pp. 137–176, doi: 10.1007/978-3-319-23847-0_5.
- [10] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?," *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 760–772, Oct. 2009, doi: 10.1016/j.jbi.2009.08.007.
- [11] J. B. Echouffo-Tcheugui and A. P. Kengne, "Risk models to predict chronic kidney disease and its progression: a systematic review," *PLoS Medicine*, vol. 9, no. 11, pp. 1–18, Nov. 2012, doi: 10.1371/journal.pmed.1001344.
- [12] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2014, pp. 85–94, doi: 10.1145/2623330.2623754.
- [13] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, vol. 2, pp. 59–78, Jan. 2006, doi: 10.1177/117693510600200030.
- [14] R. M. Conroy *et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, 2003, doi: 10.1016/S0195-668X(03)00114-3.
- [15] 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, Dec. 2011, doi: 10.1186/1472-6947-11-51.
- [16] Y. Halpern, S. Horng, L. A. Nathanson, N. I. Shapiro, B. Israel, and D. Sontag, "A comparison of dimensionality reduction techniques for unstructured clinical text," in *ICML 2012 Workshop on Clinical Data Analysis*, 2012, pp. 1–8.
- [17] L. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," in *AMIA Annual Symposium Proceedings*, 2012, pp. 505–511.
- [18] E. Dritsas and M. Trigka, "Machine learning techniques for chronic kidney disease risk prediction," *Big Data and Cognitive Computing*, vol. 6, no. 3, pp. 1–15, Sep. 2022, doi: 10.3390/bdcc6030098.
- [19] W. Luo, Q.-D. Phung, T. V. Nguyen, T. Tran, and S. Venkatesh, "Speed up health research through topic modeling of coded clinical data," in *IAPR 2014: Proceedings of 2nd International Workshop on Pattern Recognition for Healthcare Analytics*, 2014, pp. 1–4.
- [20] A. Perotte, N. Bartlett, N. Elhadad, and F. Wood, "Hierarchically supervised latent Dirichlet allocation," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, 2011, pp. 2609–2617.
- [21] J. Sun *et al.*, "Combining knowledge and data driven insights for identifying risk factors using electronic health records," in *AMIA Annual Symposium Proceedings*, 2012, pp. 901–910.
- [22] A. Singh, G. Nadkarni, J. Guttag, and E. Bottinger, "Leveraging hierarchy in medical codes for predictive modeling," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, Sep. 2014, pp. 96–103, doi: 10.1145/2649387.2649407.
- [23] F.-C. Tsui, M. M. Wagner, V. Dato, and C. C. Chang, "Value of ICD-9 coded chief complaints for detection of epidemics," in *AMIA Annual Symposium Proceedings*, 2001, pp. 711–715.
- [24] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to CARE: a collaborative engine for practical disease prediction," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 388–415, May 2010, doi: 10.1007/s10618-009-0156-z.

BIOGRAPHIES OF AUTHORS






Dr. Munusamy Chitra    is an Assistant Professor in the Department of Computer Applications, Perunthalaivar Kamarajar Arts College, Madagadipet, Puducherry. Her areas of interests include VANET, theory of computation, IoT, and machine learning. She received her B.Sc. and M.Sc. in Computer Science. She received her Ph.D. (full time) in Computer Science and Engineering from Pondicherry University. She has to her credit a number of research papers in international journals and conferences. She is the recipient of UGC-JRF award from UGC in the year June 2009, New Delhi. She can be contacted at email: chitra.jrf@gmail.com.






Abdul Kuthus Parveen    received her Bachelor Degree in Computer Applications from Perunthalaivar Kamarajar Arts College, Madagadipet, Puducherry, Pondicherry University. She has completed shorthand to her credit. Currently she is working as Police Constable in Tamilnadu Police Department. She can be contacted at email: banusagi45@gmail.com.






Murugadoss Elavarasi    received her Bachelor Degree in Computer Applications from Perunthalaivar Kamarajar Arts College, Madagadipet, Puducherry, Pondicherry University. She has completed typewriting lower to her credit. She can be contacted at email: elavarasielava72@gmail.com.



Jayamoorthy Sangeetha    received her Bachelor Degree in Computer Applications from Perunthalaivar Kamarajar Arts College, Madagadipet, Puducherry, Pondicherry University. She has completed typewriting lower to her credit. She can be contacted at email: Sangeethakavitha555@gmail.com.



Ramalingam Vaithilingame    is currently working as Assistant Professor of Computer Science in Perunthalaivar Kamarajar Arts College, Puducherry, since July 2017. He worked as Assistant Professor in Indira Gandhi College of Arts and Science from 2009 to June 2017. He received his MCA degree from Pondicherry University and M.Phil. in Computer Science from Manonmaniam Sundranar University, Tamil Nadu. His areas of interest include data structures, computer algorithms, operating systems. He can be contacted at email: sudhavaithi@gmail.com.