# Survey on Kidney Disease Prediction by using Machine Learning Technique

RavindraYadav
Assistant Professor IT Department
IET DAVV

Jay singh
Assistant Professor CSE Department
IET DAVV

**Abstract:** As per the records of WHO 19.5 million people die every year. By the year 2050 it will rise up to 65 million [1]. The medical professionals associated with diabetic diseases have some limitations, they can predict the chances of kidney faliure with the accuracy of 67% [2], for more accurate predictions doctors require a support system. The precision in predictions of kidney failure can be achieved by deep and an algorithm of machine learning . In this paper there is a lot of information about the state of art methods in deep learning and machine learning. To assist new researcher's active in this area an analytical comparison has been provided.

Keywords: Machine learning, kidney faliure, Decision Tree, Naive Bayes, Neural Network, Deep Learning and SVM.

## I. INTRODUCTION

The diseases of diabetic have created a lot of grave distress among researchers; one of the chief concerns in kidney diseases is the accurate detection and the presence of this in a human being. The techniques of early age have not been much competent in detecting it, even medical professors are less efficient in predicting the kidney diseases [3]. There a number of medical instruments available in the market but there are two main drawbacks; a). the instruments are very high-priced, and b). They are not efficient enough to calculate the kidney diseases. As per the figures of latest survey by WHO, the medical professionals are only capable to predict 67% of kidney diseases [2]. Therefore, scope and scale of research in this area is very high and large respectively.

The field of computer science has advanced at an incredible rate and has opened a huge number of opportunities in different areas of science and technology. The medical science is one such areas where the instruments of computer science can be utilized. The applied part of computer science varies from ocean engineering to meteorology.

Some of the main existing tools in computer science has been used by various medical sciences, for example, due to the rapid advancement in computation power the artificial intelligence is now reached the zenith of its existence. The machine learning is a tool that is wholly available because it never needs different type of algorithm for various datasets. In machine learning, its reprogrammable abilities open various horizons for medical sciences.

For accurately predicating the kidney diseases a great number of parameters and complex technicality is involved. This is a challenge faced by the medical sciences. Machine Learning can play a vital role in facing such challenges because its accuracy is impeccable. For predicating kidney diseases, this method of learning uses varied tools such like a feature vector and under varied situations its data-types.

The health hazard can be predicted by some algorithms like decision Tree, Naive Bayes, Neural Network and KNN. Each algorithms is different, for example, Naive Bayes works with probabilities in order to predicate heart ailments whereas Decision Tree provides classified reports for the same, and the margin of error in predicting heart ailment is lessened by the neural network. The old records of heart-patients are being used by these techniques to get accurate predictions of new patients. These predictions help doctors to save millions of lives.

This paper will provide information about the various horizons where the Machine-Learning-Technique can be applied. As we go further, this paper discusses about varied algorithms and their comparative parameters. It also illustrates upcoming prospects and its deep analysis.

## II. LITERATURE REVIEW

Researchers from various scientific backgrounds have contributed to develop this field. This machine-learning based prediction has always been one of the most curious research areas for science fraternity. There is a sudden rise in researchers working on the papers and materials associated with this area. Our chief goal is to provide all the state of artworks by various authors and researchers. Afrin Haider, Mohammad Shorif Uddin and Marjia Sultana [4] have illustrated that the datasets for kidney diseases are raw and highly superfluous and incoherent in nature. There is an immediate urgency to pre-process the datasets because at this current stage the high-dimensional dataset is deteriorated to a low dataset.

There are all sorts of features available in a dataset and they show the extraction of some crucial features. Reduction in the work of training the algorithm depends mainly upon a vital factor, i.e., selection of significant features. It results in reducing time complexity. To prove the effectiveness of algorithm the two vital parameters for comparison are used; a). Time, b). Accuracy. An effective approach has been proposed in[4] and it has contributed to ameliorate the accuracy, precision and has discovered a fact that SMO classifier and Bayes-Net performances are quite effective than 48, KStar and MLP. The performance is calculated by

putting algorithms on datasets which was collected from a software called WEKA then put into comparison by using cuurves and values of ROC and predictive accuracy. Various processes have their benefits and loss in work-done by M.A. Jabbar, Preeti Chandra and B.L Deekshatulu [5]. In order to achieve efficiency of higher class in a Decision Tree the optimisation of feature has been performed. By utilizing various features early detection of heart disease can be done. These sorts of approaches can also be utilized in other spheres of research. Some approaches excluding Decision-Tree which are dedicated to achieve the perfect goals of prediction of cardiac-ailment in humans is explained by Yogeshwaran Mohan et al [6] who has gathered raw EEG data from its devices and made use of it in training neural-network for pattern classifications. Input and output are the depressive and non-depressive characteristics of the layer which is cached and in which an algorithm is used for training that is scaled-conjugated gradient and it is utilized for achieving efficient results.

Some authors have achieved 95% efficiency with the help of trained neural networks. Researchers who work in the arena of SVM have watched outcome of Neural-Network and with the help of this technique, they classify and receive advanced and enhanced results. The Neural Network has the capability to work under high dimensional dataset. When, the Feature-Vectors which are multi-dimensional and non-linear come into play this method defeats every technique that is somewhere related to the techniques of Quantum-Contemporary.

We figured out some loop-holes after studying all techniques of State-of-Arts going. Some of them are as follows:
a) Due to variety of noises and difficulties in the dataset of medical science there seems a tremendous demand for strong algorithms which can reduce the noise.
b) There is a chance of enhancement in the efficiency and accuracy of detection of kidney diseases with the help of recent advancement in Deep-Learning.
c) Due to very dimensionality of medical dataset there are ergs to find such algorithms which can reduce and compress higher dimensionality results in gaining more execution time.

## III.MACHINE LEARNING ALGORITHM FOR KIDNEY DISEASE PREDICATION

Machine Learning is meant to aid the Artificial-Intelligence. It is used in in all the chief segments of application. Decision Tree is a graphical illustration of an exact decision that is used for model of predication. Nodes, roots and branching decisions are the main components of a Decision Tree. CART, ID3, CYT, C5.0 and J48 [7] are some approaches to build a tree. These have used the approaches to categorize the dataset by using J48, similarly [8] decision tree have been compared with the classification output of various algorithm. In medical sciences, when numerous parameters are involved in classifying a data set Decision Tree comes into play.
It is a machine learning algorithm which has compressive loom. The Decision Tree reflect some significant features. There are many parameters which affect the patient in a heart disease like sugar level in blood, blood pressure, the age and sex of the patient, some hereditary factors etc. The decision tree assists doctors to evidently identify the feature which affects the most. Among the whole throng they can also easily produce the features which influence too much. The importance of dataset can be witnessed in the Decision Tree which is completely based upon entropies and information. Over fitting and greedy methods are the two main drawbacks of Decision Tree. The reason that caused over fitting was partition of Decision-Tree datasets which were aligned to the axis, we can understand that it needs a great number of nodes if it has to separate a data. Based on greedy methods which leads to pretty fine tree, J48 in [7], illustrated the problem. If the approach is Dynamic for fine optimal consideration it might lead to an exponential category of tree which could be unreasonable.

## SVM - Support-Vector-Machine

The hyper-plane maximises the boundary of two classes the classification of SVM is performed. Support vectors [9] are the ones that define those hyper-planes. The Steps for Calculation of Hyper-plane are as follows:

**1.Set up training data**
**2. Set up SVM parameter**
**3. Train the SVM**
**4. Region classified by the SVM**
**5. Support vector**

There are various advantages and disadvantages of the usage of SVM dataset classification**.** By observing properties a medical dataset can be non-linear of high dimensionality. It is a well known actuality that SVM is the great choice for classification. Some of the advantages are :

1. At first, the regularisation of parameters which keep away from problems of over-fitting which, generally, is a substantial challenge in a decision tree.

2. Basically, the benefit of kernel tree is to replace expert opinions with kernel knowledge.

3. The SVM is a proficient process as it utilizes convex-optimisation-problem (COP), which means it
Does not have local minima.

4. When misclassification of dataset happens an Error Rate is put into testing which is a sort of great support. These features are useful in medical diagnosis which, ultimately, builds more proficient predication system. It also does not mean that it has all the goods in it. A coin does always have two sides. on second side, contains some very good features that eliminate the situation of Over-fitting which needs to get an optimized parameter flaw.
Sometimes, optimisation might result in errors. It might result in Over-Fitting.

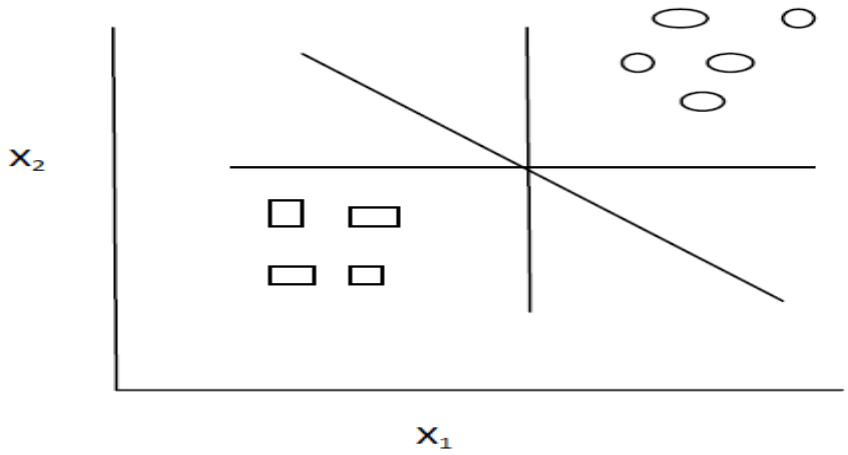


**Fig 1 SVM Classifier**

## KNN- K-Nearest Neighbour Algorithm

It is very slow and supervised algorithm. Comparatively, it takes more time to achieve trained classifications.
Like the other algorithms are divided into two steps; a). Training from data and, b). Testing it on new instances. The working principle of K-Nearest Neighbour is based upon an assignment of weight to each point of data which is known as neighbour. In this classifier, the distance is calculated for training the dataset for every K Nearest data points. Now, it is classified on the basis of widely held number of votes. Three sorts of distances are required which has to be evaluated in KNN are Minkowski, Euclidian and Manhattan.

The formula below can calculate distances [10]:
$Eucledian\ Distance = D\ x, y = (xi - yi)2ki = 1$ (1)
K = number of cluster
x , y = co-ordinate sample spaces
$Manhattan\ distance = (xi - \ )=1$ (2)
x&y are co-ordinates
Minkowski distances are generally Euclidian distance $Min = (\ - yi\ p\ )1\ p$ (3)
The sample grouping is based upon the KNN's superior class.
The appropriate grouping gives the sample deduction which is further utilized for training. Selection of the value of k plays an essential part, larger the Value of k lesser the noise and precision. The KNN algorithm is defined in the following steps:

1. D = samples, k = no. of nearest neighbours. They are used in training.
2. For each sample class create a super class.
3. For every training sample compute Euclidian distance
4. Classify the sample based on majority of class in the neighbour.

## IV.DEEP LEARNING FOR PREDICATION IN HEART DISEASE

The multi-level learning abstractions and representations are the base where the Deep-Learning is always considered as the sub-region of machine learning. The input-output layer is complemented by Multi-Processing Unit [10]. Deep learning is such an algorithm which works on the principles of feature-hierarchy. Here, the composition of lower level features forms the higher level hierarchy.

The deep learning brought the renaissance to the neural network models. There are lot many major work which is going on here by implementing stacked restricted Boltzmann Machine and auto encoder-decoder technique [11].

The researchers are impressed by the performance of this method. Performance of image processing layer wise pre-training techniques were also the areas of interest. Natural language processing and acoustic processing are the other areas of interest. For sequential feature and data, RNN is considered to be the best. Various methods are applied for the above two versions, some are LSTM which was given by Hochreiter and Schmidhuber [12]. In the sequence based task their performance is very much appreciated.

Gated Recurrent Unit (GRU) is the other modern technique of LSTM. The results of GRU are quite impressive and it's simpler that LSTM.

There is a paper [13] in which a sequential heart disease prediction has been discussed. To achieve high accuracy authors have utilized GRE. The deep learning technique is being used for medical dataset by the modern day researchers. From the serum of uric acid Lasko et al. . [14] Utilized encoder-decoder pattern. The illustration of generalised approach of deep learning is in the flow chart of Fig. 2.

There are five types of modules present in a flow chart. Every module has a specific operation. The collection of dataset from the standard repository is called Data Collection. It is then followed by a pre-processing stage where functionality in reduction of noise, and feature selection are included. The core for deep-learning is the next step because the implementation of the vital algorithmic approach adapted to manoeuvre dataset, is available. This might vary from recurrent neural network to deep belief network [15].

The data mining technique above went through an analysis of performance which became the main module as it has depicted the basic comparison of the methods explained above.

The modules, at the final discovery of virtue, will get expected results, like, the probability or percentage of the instances taking place. Here, in this scenario, it is the chances of cardiac arrest taking place in different types of patients.

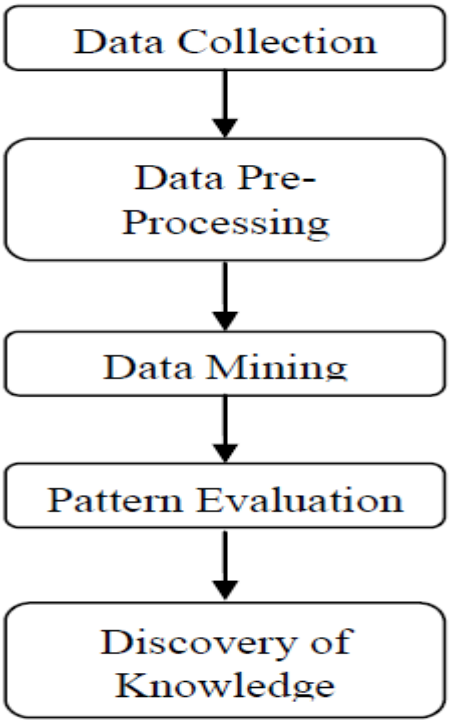


Fig 2 Flowchart of Deep Learning

## V. ANALYSIS OF AVAILABLE LEARNING ALGORITHM

Two algorithms differ in innumerable ways. So it becomes difficult to compare algorithms. The reason behind this is the complete dependency of algorithms on dataset. This results in complicating the decision making strategy, when the performance of an algorithm for an individual dataset is talked about. Only by implementing the algorithms with a particular dataset we can find out the efficiency of algorithms. The analytical comparison is required to take an appropriate decision in order to differentiate between various machine-learning algorithms. These types of works can be helpful for the researchers. In this survey we have made an effort to highlight the comparisons between various algorithms. This could benefit the beginners and new researchers to get some advantage.

| **Techniques** | **Outlier** | **Online learning** | **Over fitting and under fitting** | **Parametric** | **Accuracy** | **Execution on Technique** |
|---|---|---|---|---|---|---|
| **SVM** | It can handle outlier properly | Training on line require less Time than ANN | Performance is better than over fitting and under fitting | Non-Parametric-Mode | Always higher when compared with different parametric modes | The dataset is relatively slower |
| Decision Tree | No pivotal role played by the outlier to Interoperate the dataset by using the Decision-Tree | No online Learning is supported | It suffer over fitting and under fitting | Non parametric model | Accuracy depends on the dataset, Ensemble utilizes Decision-Tree accuracy than SVM | Needs lesser duration than other parametric modesl If there is no Over-Fitting at the same time some techniques need higher execution than decision tree |
| Naive Bayes | Hardly prune to outlier | It can perform on online testing | It does not suffer over fitting and under fitting | This is parametric | High with limited dataset | Low with limited dataset |
| ANN | More prune to outlier | Learning online take less time in ANN Than SVM | It is more prune d to over fitting than SVM | This is parametric | This is higher as compared to different Parametric-Models | Executiontime depends upon the number of layers declared and number of epochs need for testing |
| Regression which is linear | It has robust probabilistic relation and is very less prune to outlier. | For new datasets it needs training of classifier | It does not suffer from under | This is parametric | Higher for linear dataset | Less execution time required. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | fitting and over fitting | | | |

Table 1. On the basis of varied parameters there is a comparison of chief Machine-Learning Algorithm.

The Classifier of Naive Bayes explains that when there exists a high-biasness and low-variance, the training of low dataset classifier becomes effortless and is full of advantages in comparison with the classifier that has low-biasness and high-variance, like KNN. It is because the classifier, later, suffers the problem of over fitting. When the little datasets are trained they convert very quickly to needless data and time. But as we are aware of the fact that there are two aspects of a single coin, if the size of the data starts growing, there are chances of asymptomatic errors whereas the algorithm that has very little biasness and less amount of variance are very tough to ignore these types of problems.

There are some other main disadvantages of Naive Bayes algorithm, for example, it cannot learn interaction among various features. In contrary, if the regressive model of logistics considers taking care of associated characteristics not like Naive Bayes, Logistic Regression tends to provide a certain probability of mathematics. But in cases where the type of data is non-linear the logistic regression model fails to give any result. Therefore, the model-dataset feed it requires a feature or characteristic modulation which is very teasing. But it has always been friendly to the users if they want to revitalize the mode by updating characteristics within the dataset which is on the same line, though the nouvelle columns and rows arrive on the exact time, i.e., it executes well with temporal and online dataset. The Internal and external architecture of the models can be easily explained if the substantial compressibility is the major feature in the Decision Tree which is a non-parametric machine learning algorithm. There are some unfortunate drawbacks that Decision-Tree possesses like, online learning is not supported and it suffers from the over-fitting of the dataset. But there are some techniques such as J48 model which avoids over-fitting. Random forest is a an ensemble technique[16] which provides a few impugned in a decision tree, e.g., accuracy, pruning and solves the problems of an imbalanced dataset. The random forest is believed to have the capability to reorganize the perfect modes of machine learning algorithms but there is a drawback, it seizes the compressible property of the decision tree. It is considered that that Neural network and SVM are the two main competitive machine learning algorithms. But they are, actually, very different from each other with the similar motive of classification or regression. These two are non-linear classification techniques. From the derivation of statics and algebra we get SVM which construct hyper-plane in a line which is separable by N dimensional plane in order disorientate the classifiers which has large margins. It is theoretically considered that SVM provides high level of precision to all datasets. The ANN is also one the non-linear models which has plenty of drawbacks among which one is that ANN converge on each local minima. Generally, SVM never entertains these dilemmas and converge on minima of Unique and Global. The SVM can represent geometrically as it comes from a fine mathematical background. The illustration of ANN is no match to the SVM model because the complexity of ANN depends a lot upon dimensionality whereas SVM is devoid of these problems.

SVM is a unique phenomena with varied limitations. It is memory exhaustive, so the tuning becomes very difficult. The NLP faces challenges with regard to SVM as it is hard to train NLP methods. This is due to the exponential increase in complexity of duration, at the same time, we can get linear results from the ANN models.

## VI. CONCLUSION

The heart attack cases are increasing rapidly and have become a major concern in the human society. Some techniques are discussed in this paper like the State Of The Arts and others. These are utilized to predicate the cardiac anomalies. Deep learning and artificial intelligence has shown some incredible results in different areas of medical diagnosis with high accuracy. This domain is currently awaiting its chance to perform and predict. With different Machine Learning Algorithm some processes of deep learning have been considered that are useful in cardiac ailments' prediction. For finding out the best medical dataset algorithm an analysis took place. The chief targets in the upcoming days will be the ascension of the Temporal Medical Dataset where dataset vary in accordance with duration and re-training.

## REFERENCES

[1] William Carroll; G. Edward Miller, “Disease among Elderly Americans : Estimates for the US civilian non institutionalized population, 2010,” Med. Expend. Panel Surv., no. June, pp. 1–8, 2013.
[2] V. Kirubha and S. M. Priya, “Survey on Data Mining Algorithms in Disease Prediction,” vol. 38,no. 3, pp. 124–128, 2016.
[3] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu,“Prediction of risk score for heart disease using associative classification and hybrid feature subset selection,” Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.
[4] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016, 2017.

[5] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.
[6] S. Kumra, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," pp. 140–145, 2009.
[7] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.
[8] P. Sharma and A. P. R. Bhartiya, "Implementation of Decision Tree Algorithm to Analysis the Performance," Int. J. Adv. Res. Comput. Commun. Eng., vol. 1, no. 10, pp. 861–864, 2012.
[9] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J. Theor. Appl. Inf. Technol., 2009.
[10] N. Bhatia and C. Author, "Survey of Nearest Neighbor Techniques," IJCSIS) Int. J. Comput. Sci. Inf. Secur., vol. 8, no. 2, pp. 302–305, 2010
[11] J. Schmidhuber, "Deep Learning in neural networks:An overview," 2015.
[12] S. Hochreiter and J. Urgen Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[13] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques,"2008 IEEE/ACS Int. Conf. Comput. Syst. Appl., pp. 108–115, 2008.
[14] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, nd Irregular Clinical Data," PLoS One, vol. 8, no. 6, 2013.
[15] Yuming Hua, Junhai Guo, and Hua Zhao, "Deep Belief Networks and deep learning," Proc. 2015 Int. Conf. Intell. Comput. Internet Things, pp. 1–4, 2015.
[16] P. De, "Modified Random Forest Approach for Resource Allocation in 5G Network," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 11, pp. 405–413, 2016.
[17] Ashish Sharma ,Dinesh Bhuriya ,Upendra Singh, "Survey Of Stock Market Prediction Using Machine Learning Approach" , Electronics, Communication and Aerospace Technology (ICECA), 2017International conference of IEEE , 20-22 April 2017 ,pp.1-5.
[18] Dinesh Bhuriya ,Girish Kaushal ,Ashish Sharma," Stock Market Predication Using A Linear Regression ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-4.
[19] Rohit Verma ,Pkumar Choure ,Upendra Singh , "Neural Networks Through Stock Market Data Prediction" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-6.
[20] Sonal Sable ,Ankita Porwal ,Upendra Singh , "Stock Price Prediction Using Genetic Algorithms And Evolution Strategies ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.
[21] Vineeta Prakaulya ,Roopesh Sharma ,Upendra Singh, "Railway Passenger Forecasting Using Time Series Decomposition Model ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.
[22] Yashika Mathur ,Pritesh Jain ,Upendra Singh, "Foremost Section Study And Kernel Support Vector Machine Through Brain Images Classifier ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-4
[23] Pooja Kewat , Roopesh Sharma , Upendra Singh , Ravikant Itare, "Support Vector Machines Through Financial Time Series Forecasting ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-7.