End-Stage Renal Disease Risk Prediction via Machine Learning and IoT

Nipa Sarkar[1], Asha Rani Borah[2]
Department of Computer Science[1,2], New Horizon College of Engineering[1,2]
Bangalore-560103, Karnataka, India
Email: nipas94@gmail.com[1], asha.borah@gmail.com[2]

Abstract- Machine Learning Application in the field of ESR Disease Risk Prediction or ESRD (End-Stage Renal Disease) risk prediction requires various connected objects and these are the main keys for various intelligent structures, for illustration, direct access to the information about physiological values and accumulating various information about human physiology. This paper intents to give a clear idea in developing a non-invasive approach that predict various risks of dialysis patient in ESR Disease (or ESRD). This paper also represents overall synopsis about literature survey and machine learning algorithms required on building a model for gathering various data under the data analytics environment which can predict the occurrence of ESRD disease upfront by using the ESRD data generated from various IOT Sensors.

Index Terms- ESSRD; IOT sensors; ML algorithms; Artificial Intelligence.

1. INTRODUCTION

Though successful improvement is recognized from past eras, dialysis is still a complex process. Dialysis treatment is difficult to conduct and challenging due to the inaccessibility and isolation of care & treatment structures. A dialysis patient is no longer allowed to do full-time professional work and cannot be permitted in travelling. These limitations lead to a difficult life. Treatment of dialysis can be carried out at various hemodialysis centers or at home which is still unidentified and poorly proposed. The existing dialysis regulator is restricted to patient medical office appointments. Due to this constraint patient's health condition gets worsen and deteriorate with isolated appointments in public health facilities. The progression of tele-transmission methods has improved patient assurance.

Implementing dependable monitoring devices decreases the threat of complications. Progress on the Dialysis at home technique is yet to develop. Various research techniques have been implemented by using modern technological revolution to improve the dedicated medical solutions. Several problems have been recognized during the treatment of an individual on dialysis such as hypertension, hypotension, heart problems, fluid overload etc. Therefore, home dialysis can cause various additional risks, which may lead to hospitalization and death of the patient. For these reasons, ESRD patients who are suitably independent can get benefitted from the latest treatment on the agreement of the nephrologist and the paramedical crew. ESRD disease famously known End stage renal disease is end stage kidney disease. It occurs when the kidney functioning effectiveness decreases drastically. The kidney system with in the body will not be able to function as it is expected to function. The kidney filters will exhibit the abnormal behavior. In general, the kidneys filter unwanted materials and excess elements from your blood which are released through urine. The dysfunction of the kidney system can result in hazardous quantity of liquid and electrolytes deposition in our body. Recent technology proposes a better-quality technique known as Do-It-Yourself solution. This solution merges the ubiquitous technology with doctor assistance. This technique depends on collaborating patient and doctor to make health care economical and effortlessly available. Latest effective methods to care for patient at ESRD is to implement Machine Learning Techniques and IoT. The major intention of implementing these technologies are to provide caregiving environment and expand the superiority of patient's life by making use of highly efficient methods of treatment technique. Being continuously controlled by the latest technology, patient can achieve satisfactory quality of clinical treatment under the home environment itself. we are discussing about designing a smart home care architecture created by using machine learning models and IoT. This structure provides a support for the amenity offered and enables deployment. The recipients of this system are primarily ESRD patients, caregivers, nephrologist physicians, paramedical teams. Introduction of the Biomarkers represent a major technological growth in medical industry. Biomarkers are potentially suitable at every stage of medical disease discovery states and drug monitoring, which includes screening, diagnosis, risk discovery/prognosis, monitoring and so on. Measurement on these biomarkers are- saliva, plasma, urine, tissue samples, serum, and many others. The IoT and machine learning components pose various challenges in collecting fine-grained, detailed information from these biomarkers. The objective of this paper is to Introduce a survey on ongoing research to develop non-offensive methods that can predict various associated risks for ESRD dialysis patient under a smart home care environment system based on IoT. Risk is nothing but the probability of occurring a complication. Risk prediction accuracy is the frequency of predicting the right state of the patient.

Effective risk prediction may achieve by patient surveillance and predicting biomarkers values to ensure that it has normal value and the patient is out of risk. The upcoming section of this paper discusses about the literature survey and various methodology for ESRD risk prediction of dialysis at home and latest Machine Learning, IOT technologies used for it.

2. LITERATURE SURVEY

To discuss about the present status of the research associated with ESRD risk prediction, two main fields are taken into consideration. These are- (1) dialysis and (2) machine learning algorithms. Both selection embraces associated IoT biomarkers. An overall representation of the existing technology is discussed by the following subcategories which is related with this case study.

2.1 Dialysis Biomarkers

Biomarkers are commonly known as biological markers which is used to make available suitable information for analyzing the risk of a patient. The National Health Institutes defines biomarker as a distinguishing measurable indicator using which a particular biological processes, pathogenic processes, pharmacologic responses, severity or existence. The kidneys of human body preserve constituency and amount of liquids in the body via regulatory system and balances water level, electrolytes, acidity etc. and controls the elimination of contaminants and unsolidified. GFR or glomerular purification percentage and albuminuria are the common methods used to check the kidney functionality. BIA or the Bio Impedance Analysis techniques are widely used to maintain the status of consistency in patients diagnosed with dialysis. YKL-40 is useful alternative inflammatory biomarker for dialysis patients Which can rise in plasma through diminishing renal function irrespective of soreness.

Latest developments in the field of molecular biology give rise to capable biomarkers designed for AKI as well as CKD detectors nevertheless, further research and development is required to contrivance these effectively for the medical practice to simplify primary diagnosis and observation on the disease evolution. Some of the best interpreters of dialysis are - cystatin C, IL-18, Urinary NGAL etc. To understand and choose the best IoT sensor for dialysis prediction, survey is done on the impact factor of biomarker is observed on various dialysis patient’s health and several aspects is been noted which influence biomarkers. For illustration, cystatin C is a biomarker used for kidney functionality that reduces the temperature of the body when examined using mice. Cystatin C can deliver innovative perceptions in addition with ESRD & hypertension.

2.2 Machine Learning Algorithms

Machine learning is a domain that enables the system automatically to learn through experiences and improve performance by itself without being explicitly programmed by a developer. The main aim of Machine learning technique is the expansion and advancement of system programs or the algorithms which is able to access various types of data and learn and experience from it for assuming the upcoming occurrence. The main intention of machine learning is to allow the system/computers experience automatically without human involvement or support and perform actions consequently. Machine learning models are categorized as supervised or unsupervised. Supervised ML algorithm techniques can implement past learnings from historical data to new data and by doing so it can predict the result of future occurrences. The system is able to provide targets for any new input after sufficient training.

Whereas, unsupervised ML process are used for the training dataset which are not classified as well as not labelled. Unsupervised learning system studies the inference of system function to define a concealed assembly of unlabeled data

Semi-supervised ML algorithms stand in middle of supervised and unsupervised technique of learning which include some amount of labelled and huge quantity of unlabeled data. System that follows semi-supervised technique can significantly progress learning precision. Semi-supervised technique is selected when assimilated labelled data involves trained and appropriate possessions in order to learn from it. Otherwise, acquiring unlabeled data usually doesn’t need supplementary resources.

Reinforcement machine learning is also a ML algorithm that interacts with its surrounding environment by performing some actions and determines errors or rewards based on learning experience. Trial and error search, delayed reward are important features of a reinforcement learning ML model. The reinforcement method maximizes the performance by allowing agents to auto determine the suitable behavior of any specific situation.

Machine learning allows analysis of enormous extents of data. It provides faster, and more precise results to categorize profitable prospects or risky aspects Advantages of the ML algorithms are- (1) Machine learning algorithms help to track real time behaviors that can be used as inputs for companies to gain better results and reduce operation cost. (2) Uncertainty can be determined by using machine learning techniques along with certain confidence up front which leads to reduction of chaos. Here, we will be discussing various machine learning algorithms which can help to predict the early stage renal diseases. the ensemble classifier will help in predicting the disease early stage. This will help the patients in taking counteractive measures upfront and reduce the chances of kidney failures.

3. PROPOSED FEATURES OF THE SYSTEM

The ESRD risk prediction system has three supervised Machine Learning algorithms to classify the software liabilities based on past and historic data. The ML classifiers are- 1) Naïve Bayes or NB, 2) Decision Tree or DT, 3) Artificial Neural Networks or ANNs. The valuation procedure determines that ML algorithms can be applied efficiently with high accuracy percentage and better performance. Propose System involved the below process flow in sequential order:

(1) Defining the problem: Identifying the right factors required for bug prediction.
(2) Dataset Labels creation: Converting the base dataset to analytical data set.
(3) Base Models: implementing the baseline models (Logistic Regression).
(4) Benchmarking Models: Implementing benchmarking models (Random Forest, Decision trees, Naïve Bayes, ANN).
(5) K Fold Validation: Validating the algorithms across the datasets.
(6) Accuracy Metrics: comparing the accuracy metrics.
(7) Model Publishing: Deploying the model.

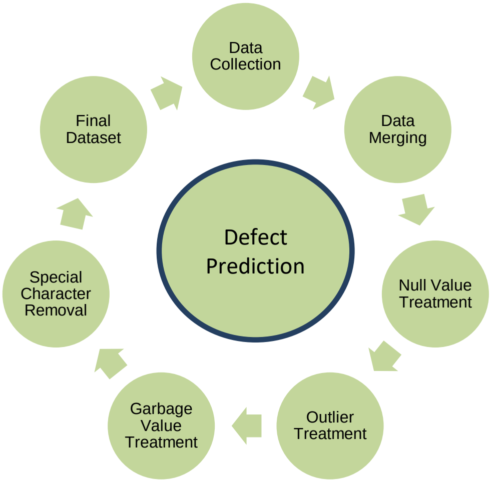

Fig. 1. Machine Learning Model Flow

As the ensemble classifiers are used, the model becomes a better choice for predictions. Validating the system through K- Fold cross validation technique results in effectively working of the model for any kind of data.

4. SYSTEM DETAILS

The procedure of designing and defining the architecture of a system, it's components & modules and available data to satisfy its requirements is termed as system design. System design describes about the models used for analysis, architecture required to build the models and the statistical concepts behind all the algorithms used.

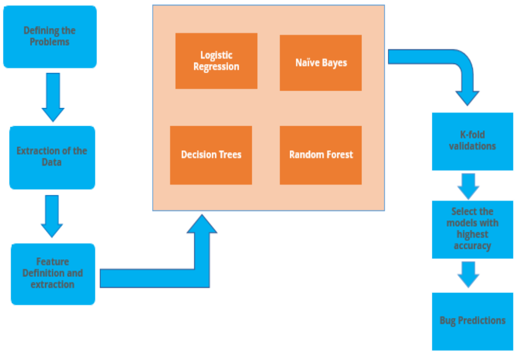

Fig. 2. System Architecture

The above system structure attempts to resolve the risk issues of ESRD. Solving the problem starts with identifying the right factors required for bug prediction. The prioritized factors are considered for creating analytical data set. Base line models and benchmarking models are implemented on top of analytical dataset. All algorithms are validated through k fold validation methodology to find the right accuracy. Any supervised ML algorithms will require a systematic flow of the below-mentioned steps. These are generalized frameworks which can help in defining the problem better and executing the all the phases of the project in structured manner.

4.1 Problem solving framework

The main important characteristics of the any machine learning model is defining the problem better. The key factors required for the analysis are identified in the process. It will help in listing down the factors without bias. Steps Involved are:

(1) Defining the problem: Define the current state and need for doing the problem.

(2) Factor Map: List down all the factors required for analyzing the problems.

(3) Hypothesis Generation: Generated hypothesis associated with the problems statement.

(4) Prioritization Matrix: Identify the factors that are based on actionability and feasibility matrix which leads to prioritize important factors.

(5) Important Factors: Prioritization matrix will help in analyzing the problems associated with the matrix.

4.2 Data preprocessing (Exploratory Data Analysis or EDA)

Majority of the time the data collected for running machine learning algorithms is not available readymade. Hence, the data requires pre-processing which will help to get the right predictions. The data collection is the most vital and significant steps among all the steps of machine learning algorithm model predictions.

4.3 Baseline model

Logistic regression is the baseline model which is widely used in most of the projects. As the bug prediction is a binary response variable, the bug behavior can be predicted by multiple algorithms. Logistic Regression is considered as base model and other algorithms are measured as benchmarking models under the logistic regression algorithm. Steps in Logistic Regression:

(1) Analytical dataset creation: Create the base data set required for the analysis.
(2) Train test split: Dividing the dataset in train and test for executing ML algorithms.
(3) Baseline model's creation: Construct the baseline models with the given set of variables.
(4) Validating the performance of the models: Identify the Accuracy score, Confusion Matrix, Roc- AUC curve designing, Probability curve, Improving model performance and Defect Detection models.

4.4 Benchmarking models

Random Forest Classification OR Bagging technique is used to decrease the variance of estimation by merging the result of numerous classifiers which are modelled on diverse sub-samples on similar datasets. Higher quantity of models provides parallel result (better performance) than lower number of models. Under some assumptions variance of collective predictions is reduced to 1/n (n: number of classifiers) of the original variance.

4.5 Model validation

Confusion matrix is commonly used as error matrix. Confusion matrix is an explicit table outline that permits visualization of the performance & presentation of algorithm. Each row of the matrix signifies the instances in a predicted class while each column represents the instances in an actual class [2].

ROC curve is a receiver operating characteristic curve. It determines a graphical plot that demonstrates the analytical capability of a binary classifier structure as its insight threshold is wide-ranging.

4.6 cross validation

Cross Validation technique is very useful method for evaluating the outcomes of machine learning techniques. It helps in knowing how the ML model would simplify an automated data set. Cross Validation technique is used to evaluate the accuracy of the predictions by the model in practice. When a ML problem is provided, two type of data sets are generated— known data (training data set) and unknown data (test data set). By using cross validation, machine learning model is tested during "training" stage to validate for overfitting and to predict the performance of ML model while generalizing independent data, which is the test data set given in the problem.

K-Fold Cross Validation: K-Fold Cross Validation technique is widely used in machine learning. The overall steps to perform K-Fold technique are listed below:

(1) Shuffle the dataset randomly.
(2) Split the dataset into k groups
(3) For each unique group:
(4) Take the group as a hold out or test data set
(5) Take the remaining groups as a training data set
(6) Fit a model on the training set and evaluate it on the test set

(7) Retain the evaluation score and discard the model
(8) Summarize the skill of the model using the sample of model evaluation scores

4.7 Algorithm overview

4.7.1 Logistic regression

Logistic regression algorithm of ML is a method of estimation, like the OLS regression. Typical application areas are cases where one wishes to predict the likelihood of an entity belonging to one group or another. Logistic Regression fits an S-shaped curve to the data.

4.7.2 Decision tree & random forest:

Decision tree is also a supervised learning model with pre-defined target variable commonly used in classification of the problems. Decision tree works with the variables which has categorical and continuous input or output **[1]**. In this method, the inhabitants are grouped into several homogeneous sets depending on most substantial differentiator in input variables. The common terms applied along Decision trees are:

- Root Node: It characterizes the whole inhabitants or sample which can further be grouped into two or more standardized sets
- Splitting: procedure of dividing a node into two or more sub-nodes [1].
- Decision Node: If a sub-node is further spliced into further sub-nodes, then it is termed as a decision node [1].
- Leaf or Terminal Node: The nodes which cannot split further is called Leaf or Terminal node [1].
- Remove sub-nodes: Removing sub-nodes from a decision node is termed as pruning. It is the opposite process of splitting [1].
- Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree [1].
- Parent and Child Node: Node which is divided into sub-nodes is called parent node of sub-nodes. Sub-nodes are the child of parent node [1].

4.7.3 Naïve Bayes

Naïve Bayes algorithm uses the classification technique grounded on Bayes' theorem along with an assumption of predictors [3]. Naive Bayes method predicts the occurrence of a character inside a class which is unconnected to the occurrence of any further character. Naive Bayes system model is simple to build and predominantly suitable for very huge data sets. Along with effortlessness and simplicity, Naive Bayes is well known for outperforming highly erudite classification methods [4].

4.7.4 Artificial Neural Networks

Deep learning is also commonly termed hierarchical learning or as deep structured learning. Deep learning is a part of machine learning approaches which is mainly created on learning data illustrations and to specify the task-specific algorithms [4]. Deep learning technique is categorized into supervised, semi-supervised and unsupervised. Deep neural networks, deep belief networks and recurrent neural networks are most important Deep learning architectures which is been introduced to various fields including computer vision, NLU (Natural Language Understanding) speech(or voice) recognition, NLP(Natural Language Processing),audio recognition, social network filtering, machine translation, bioinformatics, drug design and board game programs [6]. In all these above-mentioned fields deep learning has generated comparable positive results, in some cases, it is even superior to human specialists [6]. Deep learning techniques are vaguely inspired by data processing and communication forms in nervous systems yet have various differences from the structural and functional properties of biological brains (especially human brain), which make them incompatible with neuroscience evidences.

5. CONCLUSION

Machine Learning Application in End-Stage Renal Disease Risk Prediction or ESSRD risk prediction requires various connected objects and these are the main keys for various intelligent structures for illustration, direct access to the information about the corporal and physiological values and accumulating various information about human physiology. This paper intents to develop a non-invasive approach that predict various risks of dialysis patient in End-Stage Renal Disease (ESRD). In this paper, we propose various machine learning techniques to build a model for gathering various data under the data analytics environment which can predict the occurrence of ESRD disease upfront by using the ESRD data generated from various IOT Sensors.

REFERENCES

[1] Dileep Kumar G "Chapter-1 Tree based modelling techniques: IGI GLOBAL,2019.

[2] Dejun Mu, Junhong Duan, Xiaoyu Li,Hang Dai, Xiayan Cai, Latin Guo. "Expede Herculem: Learning Multi Labels from Single Label", IEEE Acess, 2018.

[3] Roy, Sandip Kumar, and Preeta Sharan. "Application of Machine Learning For Real-time Evaluation of Salinity (or TDS) in Drinking Water Using Photonic Sensor", Drinking Water Engineering and Science Discussions, 2016.

[4] Han Feng. "The application of artificial intelligence in electrical automation control", Journal of Physics: Conference Series, 2018.

[5] The International Conference on Advanced Machine Learning Technologies and Applications(AMLTA2018)", Springer Nature, 2018.

[6] Iman Raeesi Vanani, Morteza Amirhosseini. “Chapter 3 Deep Learning for Opinion Mining”, IGI Global, 2019.