A Predictive Analysis of Chronic Kidney Disease Using Machine Learning

by

MD.SHAFAYET KHAN
17301093
NAZIHAN AFRIDA
18301090
MUNIA RAHMAN
19101523
SUJANA ISLAM
19101152
ANANYA BANIK
21101342

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2022

Declaration

Hereby it is proclaimed that

1. We created the thesis from scratch while earning our degree at Brac University.
2. The thesis does not include any already published or written by a third party content, unless it is properly referenced in the references.
3. The thesis does not include any content that has already been approved or submitted for consideration for another degree or certificate at a university or other institution.
4. All significant sources of assistance have been recognized.

Student's Full Name & Signature:

Shafayet Khan

MD.SHAFAYET KHAN

17301093

NAZIHAN AFRIDA

18301090

Munia Rahman

MUNIA RAHMAN

19101523

Sujana Islam

SUJANA ISLAM

19101152

Ananya

ANANYA BANIK

21101342

Approval

The thesis/project titled ”A Predictive Analysis of Chronic Kidney Disease Using Machine Learning” submitted by

1. MD.SHAFAYET KHAN (17301093)
2. NAZIHAN AFRIDA (18301090)
3. MUNIA RAHMAN(19101523)
4. SUJANA ISLAM(19101152)
5. ANANYA BANIK(21101342)

On September 20th, 2022, the requirements for the B.Sc. in Computer Science degree have been acknowledged as partially satisfied in a suitable manner. of Summer, 2022.

Examining Committee:

Supervisor:
(Rafeed rahman)

Rafeed Rahman

Lecturer
CSE Department
Brac University

Co-supervisor:
(MD Sabbir Ahmed)

Md. Sabbir Ahmed

MD Sabbir Ahmed

Lecturer
CSE Department
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

____

Sadia Hamid Kazi

Head of the department
Department of Computer Science and Engineering
Brac University

0.1 Abstract

Chronic kidney disease (CKD) is a determined disease condition having critical grimness and death rate that influences the whole grown-up populace brought about by either renal pathology or diminished renal capabilities. Early location and powerful treatments might have the option to end or diminish the growth of this constant condition to last stage, where dialysis or kidney transplantation is the main life-saving choice for patients. In this examination, we have investigated the opportunities for early chronic kidney disease expectation utilizing an assortment of machine learning algorithms. Here, a reasonable CKD dataset was taken from Tawam Clinic in AlAin city (Abu Dhabi, Joined Middle Easterner Emirates). We have proposed Support vector machine (SVM), Random forest algorithms (RF), Logistic regression (LR), Multinomial naive bayes (MNB), LSTM and contrasted their results with figure out the best exactness among the models. As a result, the models yielded outstandingly great order precision, with a LSTM exactness of 0.95 percent. The result of the review shows that improvements in machine learning (ML), with the assistance of prescient knowledge, comprise a reasonable climate for recognizing commonsense arrangements, which thus exhibit the prescient capacity in the space of renal illness and then some.

Keywords: Machine Learning; Prediction;logistic regression;MNV;Random Forest; SVM; LSTM.

0.2 Acknowledgement

First and foremost, praise is due to the All-Powerful Allah, with whose assistance we were able to complete our research without too many obstacles. Second, we want to express our gratitude to Rafeed Rahman Sir, our manager, for his considerate help and advice. He helped us whenever we needed it. And eventually, it might not be possible if our parents don't continue to support us. Thanks to their kind prayers and encouragement, we are currently getting ready to graduate

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Introduction

Practically we all have grown up with the saying that "Health is wealth" since adolescence. Wellbeing is a realm of whole physical, scholarly and social prosperity. There are numerous infections, vices which make individuals debilitated and those can be deadly as well. Among them chronic kidney disease (CKD) is known as silent killer[25]. Most of the time, CKD creates no side effects and is possibly found at it's advanced stage [25].This is one reason why a large number of individuals pass on from it and at times the genuine explanation of their demise stays obscure to all. Our kidneys have a huge number of little veins that behave like channels, called nephrons. It is the utilitarian unit of our kidneys. It eliminates side-effects from our blood. On the off chance that in some way it comes up short and the kidneys lose the capacity to channel byproducts, it can prompt kidney disappointment now and again [12]. Specialists allude to any irregularities of the kidneys as kidney sickness, regardless of whether the mischief is minor. A 'constant' disease is one that doesn't improve completely and goes on for a drawn out timeframe. 'Serious' doesn't infer persistent. CKD can be brought about by an assortment of elements, the most predominant of which are diabetes, hypertension. Since just a small bunch of the reasons of CKD are absolutely treatable, intensive testing to find a reason is regularly superfluous insofar as blood tests uncover that kidney work is steady. A sweep of the kidneys will be led in the event that somebody has fundamentally lessened kidney work, disintegrating renal capacity, or related issues like kidney uneasiness [28]. Early distinguishing proof and therapy can regularly forestall the movement of CKD. At the point when kidney sickness proceeds, it can prompt renal disappointment, which requires dialysis or a kidney relocate to remain alive[32]. Many individuals can't understand that, they have such kidney illnesses until it's high level. It's ramifications can be lethal. In any case, in the event that we can foster an early expectation model for CKD with a more noteworthy exactness it will be extraordinary and it can save individuals from CKD or possibly delayed down the deadly outcomes of CKD [26]. To foster such model, we can take the assistance of profound learning, ML algorithms like Linear regression, Logistic regression, Decision trees, Naive Bayes, SVM, kNN, K-Means, Random forest and that's just the beginning.Beside those, we especially want to see the performance of Long Short Term Memory (LSTM) in the prediction of CKD. If it works well and we can find a good accuracy then it will be a unique achievement in the prediction of our research field. Afterward, we will likewise attempt to see which calculation is the most appropriate for the model.

1.2 Research overview

We have picked the topic "A predictive analysis of chronic kidney disease using machine learning "on the grounds that the expectation framework for CKD will assist specialists with anticipating kidney illness in beginning phase which will bring about saving huge number of lives. The passing rate in CKD is extremely high. In this way, foreseeing the kidney illness in beginning stage is vital. We have put to utilize a dataset and after data preprocessing we have cleaned the dataset by eliminating the unnecessary columns and checking if there is any missing values or not. From that point onward, we have separated the dataset into training and testing and implemented in various ML algorithms. Those algorithms turn over variation precision and anticipate regardless of whether the patient bears the CKD disease.

1.3 Research problems

CKD is a significant general wellbeing worry all over the planet, especially in lowand center pay countries. CKD is a condition where the kidneys quit working appropriately and can't channel blood as expected. Around 10 percent of the worldwide population has CKD[26]. A great many individuals bite the dust every year because of an absence of cheap therapy choices, with the quantity of more seasoned individuals expanding CKD has ascended as a significant reason for mortality around the world, as per the International Society of Nephrology's Global Burden Disease research, with the quantity of passing developing by 82.3 percent over the most recent twenty years [7],[4]. Moreover, the quantity of affected people with CKD is rising, requiring kidney transplantation or dialysis to save their lives. [4],[11],[3]. CKD has no side effects in its beginning phases, hence testing might be the best way to decide if the patient has renal sickness. Early recognizable proof of CKD in its beginning phases can help patients get fitting treatment and stay away from the advancement of CKD[7]. An individual with one of the CKD risk factors, like a family background of renal disappointment, hypertension, diabetes ought to be assessed consistently. There are also some common kidney diseases like acute kidney injury (AKI), kidney stones, kidney infections, kidney cysts, kidney cancer etc. But they are not as severe as CKD. Now, let us see some differences between CKD and other common kidney diseases or AKI which will help us to understand why CKD is so severe and why should we take it so seriously.

AKI happens unexpectedly and is much of the time reversible then again CKD creates over a significant stretch and it is for the most part not reversible .
AKI happens when the kidney unexpectedly bombs because of a physical issue, drug, or illness. CKD is the slow loss of kidney works that is mostly brought about by hypertension, diabetes, and a fiery condition known as glomerulonephritis .
The side effects of AKI are electrolyte lopsidedness, check in the urinary parcel, blood in the urine pr diminished urine yield. Then again, side effects of CKD may not create until next to no kidney work remains [22].

Other kidney sicknesses cause torment in the side or lower back, outrageous thirst, faintness, and powerless fast heartbeat. Going against the norm, numerous different issues might create with CKD like iron deficiency and hyperphosphatemia.
Patients who experience typical kidney sickness need brief dialysis until their kidneys mend yet CKD patients can take dialysis therapy that plays out a portion of the elements of the kidney it isn't for all time reparable[22].
With treatment other kidney sickness might get back to typical condition yet on account of a CKD patient he really wants to have dialysis until the end of his life .

1.4 Research objectives

This study attempts to create a system that can effectively and accurately diagnose CKD at an early stage. We have already discussed how deadly chronic kidney disease can be if it isn't identified early. Although the damage caused by this disease can be permanent, early discovery and treatment can help to lessen the amount of damage done as well as the fatality rate. Researchers are working on establishing such a system, and a number of effective systems have already been presented that can detect CKD with remarkable precision. However, this field requires a more precise system that would deliver the highest level of accuracy. For this we want to have a better understanding of chronic kidney disease by learning about the symptoms and effects of this condition. Our main goal is to create a model that is more efficient than current systems for the prediction of CKD by getting knowledge about current research and detecting methods. Then we will try to improve our system's performance and also make suggestions about how to improve the existing model

1.5 Challenges faced

However CKD is intimately acquainted subject yet the path of our work was not exceptionally simple by any means. From the get go, we confronted a ton of challenges to set up a unique dataset for our research. We directed a few gatherings to comprehend what our essential objective will be. After collecting the raw data, our fundamental concern was to prepare that data of model execution. Along that, picking the fitting machine learning algorithms was a difficult issue. For that, we need to become familiar with a few new cycles and procedures of ML, AI and data science. We attempted the best to conceal these issues and endeavored to appropriately finish this exploration.

1.6 Paper orientation

Our entire proposition is portrayed in six parts. From the beginning, in section 1, we attempted to give an overview of our topic. We began with a basic of CKD in worldwide viewpoint. We additionally discussed the challenges that we needed to face to finish our work. In section 2, there is a nitty gritty depiction about the current work and writing survey that are firmly connected with renal illness, renal capabilities. Then, in part 3, we needed to exhibit our work plan with a pictorial portrayal. We likewise examined about our dataset, data pre processing and feature selection method. From that point onward, in part 4, we showed their accuracies with pictorial representation and dissected the outcome with graphical portrayal. And then, we wanted to analyze our result deeply with a small comparison with other existing works. We also pointed why our work is unique. At last, in section 6, we finished up our paper.

Chapter 2

2.1 Literature review

Presently we will examine about a portion of the past works connected with our examination field
Here, in the paper [1], the creators made beginning prescient models for ongoing kidney sickness utilizing four AI strategies: order and relapse trees, calculated relapse, support vector machines, and multi-facet perceptron brain organizations. They utilized an informational index from Apollo Hospitals, India which is accessible in online at UCI AI store. They began with 24 boundaries and dispersed them into three gatherings which are gotten from blood tests, pee tests and other affecting boundaries. As they needed to make an ideal subset of boundaries they wound up with 7 most enlightening boundaries. These are hemoglobin, hypertension, glucose, explicit gravity, egg whites, discharge cells and creatinine. They have found the multi-facet perceptron brain organization and calculated relapse mutually to be most proficient models to anticipate persistent kidney infection utilizing the ideal boundaries. Furthermore, in this manner, they have tracked down a productive way to deal with foresee persistent kidney sickness[12].

In the paper-"Computational Intelligence Approaches for Prediction of Chronic Kidney Disease" [2], we can see the creator's motivation to explore the advancement of AI work processes for early expectation of CKD in light of choice emotionally supportive networks. ML grouping strategies have a critical effect and obligation to foresee kidney sickness in the persistent illness research local area. They utilized a half breed ML model to break down kidney infection. This shows the proficiency of the ML approaches in anticipating the CKD of 97.8regularly characterized as kidney harm and happens when the kidneys can't spotless the blood proficiently. That's what results show, these arrangement techniques perform well for anticipating kidney illness. In absolute five stages, stage 5 is the last phase of kidney sickness, which happens when the human kidney totally stops working. This application apparatus can identify kidney harms at all stages where kidney infection is reasonable. NB and RF characterization procedures both are extremely valuable for early discovery of renal sickness in patients. In this review, six ML methods were utilized to recognize kidney sicknesses at beginning phase. Their review inferred that the precision of DT and NB was 91power contrasted with NB. Minor issues happen in stage 2 kidney sickness and ordinarily have no side effects of kidney harm in stage 1. They utilized various ML libraries to break down this kidney dataset. Early expectation of constant kidney sickness is critical for determination and therapy. In this way, the

NB and RF grouping strategies anticipated the greatest number of renal patients[21].

The creators of the paper [3] of given reference, have examined how to have the option to answer an assortment of kidney illness inquiries by interfacing expanding sums and different methodology kidney sickness information through a strong new AI system. ML approaches can further develop logical execution while quickly growing accessible multi-level datasets. As the accessibility of far reaching renal omics frames (transcriptomics, proteomics, metabolomics, genomic sequencing) expands, ML approaches for breaking down human kidney datasets are turning out to be progressively significant. This segment depicts how to apply ML to the investigation of CKD. Altogether, they zeroed in on the most proficient method to make ML approaches ready to grasp the connection among genotype and aggregate[18].

As per the paper [4], ongoing kidney illness (CKD) is a moderate and irreversible disintegration of the design and capacity of the kidneys, particularly glomerulus filtration rate that happens in a few times, months or years. Here, the creators have expressed that their plan to apply different ML calculations to assess and look at the proficiency in the discovery of persistent kidney illness and other execution boundaries. Persistent kidney illness dataset from the University of California, Irvine AI store, was utilized and eight administered ML models were created with the assistance of Python programming. One more work was finished by Huseyin et al. where they made do by choosing the highlights of the dataset before the calculation is applied. One of the most utilized and most exact datasets in the execution of the ML calculation is UCI Record storehouse. The dataset contains 400 occasions and 25 cases quality. By assessing various models, a similar examination between eight ML models is introduced. Execution boundaries, for example, exactness, responsiveness, F1 score and so on. Arbitrary woods showed the most elevated precision with 99.75 percent[20].

As per the paper [5], ongoing kidney sickness expectation in clinical field is a moving errand to do. In this paper, the creators have referenced around 3 ML calculations which are Decision Tree (DT) calculation, Naive Bayesian (NB) calculation and so on. The general presentation of the above models are in examination with each unique with the aim to choose the great classifier in anticipating ceaseless kidney problem for given dataset. In the dataset viable, the Random woods accomplished better prescient execution regarding grouping precision, particularity, responsiveness boundaries. The proposed informational index has 25 credits, 11 numbers, and 14 nominals. A portion of the dataset credits are age, bp, sg, al, su, turf, hemo, pcr, wc, rc order. The dataset is partitioned into two gatherings: preparing and testing. The proportion to prepare dataset to test information is 70higher, and it anticipated persistent kidney illness with 99.25additionally expressed that, they will keep on further developing the precision execution of prescient frameworks in brain organizations and profound learning calculations[14].

The creators show in their ”Prescient Analytics for Chronic Kidney Disease Using Machine Learning Techniques” paper [6], some AI strategies for anticipating the ongoing kidney sickness utilizing clinical information. Four AI techniques are investigated including K-closest neighbors (KNN), support vector machine (SVM),

strategic relapse (LR), and choice tree classifiers. These models are worked from persistent kidney illness dataset and the working of these models are contrasted together all together with pick the best classifier for anticipating the ongoing kidney sickness[9].
As per the paper [7], "Ongoing Kidney Disease Prediction utilizing Machine Learning Models" the creator proposes best expectation system for CKD and calculation to foresee CKD at a beginning phase by utilizing information preprocessing , information change and different classifiers like Decision Tree, Random Forest and Support Vector Machine. The consequences of the structure show promising aftereffects of better forecast at a beginning phase of CKD. CKD is a condition wherein the kidneys are harmed and can't channel blood. In addition, they present confirmations of early distinguishing proof and care of CKD can work on the state of the patient's life[17].

In the paper [8], 'A Machine Learning Methodology for Diagnosing Chronic Kidney Disease' the creators have referenced that CKD is a worldwide medical condition with high death rate and with no conspicuous side effects in the beginning phase. Consequently, early forecast is important to battle this ailment and guarantee great treatment. As indicated by the creators, the data of dataset they utilized were acquired from University of California Irvine (UCI). Then they applied calculated relapse, arbitrary backwoods, support vector machine, k-closest neighbor, guileless Bayes classifier and feed forward brain organization to lay out models. In the wake of looking at their presentation, exactness, accuracy and review results, at long last Random woodland is decided to execute this framework and the precision acquired is 99.75 percent. Utilizing least number of highlights, the result predicts on the off chance that the individual has CKD or not[16].

As per the creators of paper [9], 'Identification of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors' they expected to test the capacity of AI calculations for the expectation of ongoing kidney illness which is a basic medical condition. They creators referenced that during highlight choice they have tracked down that hemoglobin, egg whites, and explicit gravity have the most effect on foresee the CKD. They applied Logistic relapse, support vector machines, arbitrary backwoods, and inclination helping calculations and every one of them have been prepared and tried utilizing 10-overlap crossapproval. Among them, they acquired an exactness of 99.1 as indicated by F1- measure from Gradient Boosting classifier. Hence, utilizing suitable dataset and littlest subset of highlights they had the option to effectively anticipate CKD[15].

As per the paper [10], 'An end stage kidney sickness indicator in view of a counterfeit brain networks group', not many examinations on the arrangement or conclusion of constant renal illness have been led as of late. T. Di Noia et al. distributed a product application in 2013 that utilized a counterfeit brain organization (ANN) to order persistent status and foresee the probability of end stage renal disease (ESRD). The classifiers were prepared utilizing information accumulated more than a 38-year time span at the University of Bari, and the assessment depended on accuracy, review, and F-measure. The product apparatus offered has been made available as an Android versatile application as well as an internet based web application[6].

In the paper [11], 'Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease', K. A. Padmanaban and G. Parthiban utilized AI strategies to analyze ongoing renal sickness in diabetic people. They utilized 600 clinical information from a top diabetic examination foundation in Chennai for their review. The creators utilized the WEKA program to assess the dataset utilizing the choice tree and Nave Bayes approaches for order. They found that the choice tree strategy beats the Nave Bayes calculation by 91 percent [10].

Chapter 3

3.1 Work plan

The purpose of this study is to analyze and predict CKD. Here, using machine learning model, that will process the data, we can get our expected output. The following figure illustrates the work plan with a pictorial representation.

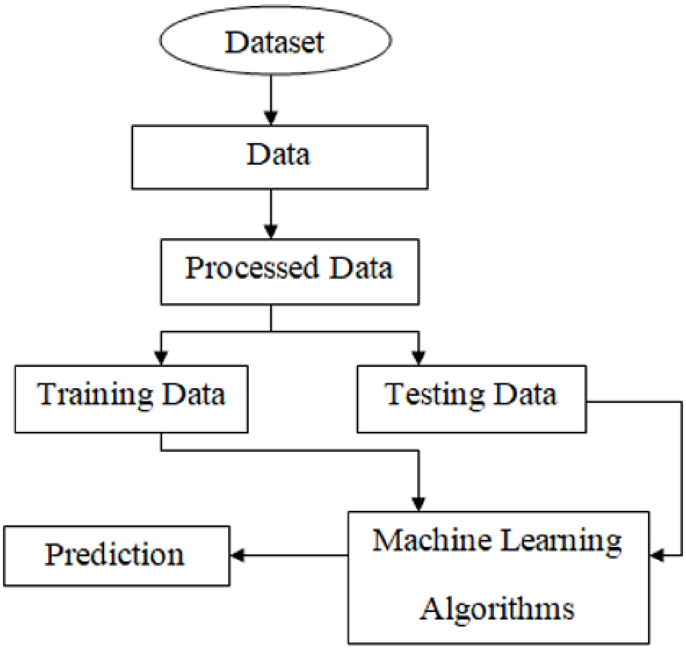


Figure 3.1: Working plan overview

3.2 Dataset

In this survey, we have examined the electronic clinical record dataset of 491 patients assembled from Tawam Clinic in AlAin city (Abu Dhabi, Joined Middle Easterner Emirates), over the timeframe from January 1 to December 31, 2008 [24], [19]. Patients included 241 women and 250 men, with a mean time of 53.2 years. The dataset contains hard and fast 22 segments and 491 lines. Beneath in the table we have examined pretty much every one of the highlights accessible in the dataset.

Column Names	Explanation
Sex	If the patient is a man (1) or women (0).
Age Baseline	Age of the patient.
History Diabetes	Patient's history of diabetes.
History CHD	Patient's history of coronary heart disease.
History Vascular	Patient's history of vascular diseases.
History Smoking	Patient's history of smoking.
History HTN	Patient's history of hypertension.
History DLD	Patient's history of dyslipidemia.
History Obesity	Patient's history of obesity.
DLD meds	Assuming the patient has taken dyslipidemia prescriptions.
DM meds	Assuming that the patient has taken diabetes drugs.
HTN meds	In the event that the patient has taken hypertension drugs..
ACEIARB	On the off chance that the patient has taken ACEI or ARB.
Cholesterol Baseline	Level of cholesterol.
Creatinine Baseline	Level of creatinine in the blood.
eGFR Baseline	Estimated glomerular filtration rate.
sBP Baseline	Systolic blood pressure.
dBP Baseline	Diastolic blood pressure.
BMI Baseline	Body-mass index of the patient.
Time To Event Months	Months from follow-up start to serious CKD occasion or last visit.
Event CKD35 (Target)	Assuming that the patient had CKD.
TIME_YEAR	Years from follow-up start to serious CKD occasion or last visit.

3.3 Data Preprocessing

Datasets are vulnerable to missing, boisterous, excess, conflicting information, particularly clinical datasets .Working with poor quality information prompts poor quality outcomes. To make it fit for the calculations we want to go through a few handling steps. We first looked for unnecessary columns in the dataset and eliminated that. Then we looked if there were any rows with missing values. But found nothing such like that.

3.4 Feature selection and Train-Test Split

Feature selection decreases the quantity of information factors while fostering a prescient model. To accomplish better accuracy for any model, it is fundamental to distinguish related elements and eliminate unimportant or less significant highlights from a bunch of information that don't add to the objective variable. For our work we have used univariate feature selection method. It works by selecting the best features based on the univariate statistical test. Compare each characteristic with the target variable to see if there is a statistically significant relationship between them. Also known as analysis of variance (ANOVA). Applying univarite selection we have chosen 10 features for our model. They are Time To Event Months, eGFR Baseline, Creatinine Baseline, Age Baseline, DM meds, History CHD, History Diabetes, ACEIARB, sBP Baseline and Event CKD35. Then we focused to separate data into two sections: training and testing. The model gains from a training dataset that contains known yields, and the training helps later partake in the speculation of different data. In this way, utilize the training dataset to prepare the model and utilize a test set that doesn't show up in the training data model for precision evaluation. We have utilized sklearn to part the dataset into training and testing and kept the default proportion of 80 percent for the training dataset and 20 percent for the test dataset. We have chosen the features X and the targets Y, keeping up with the principles.

3.5 Implemented models

This part examines about the algorithms we are utilizing in our work. First we had to understand will our output would be categorical or numerical. As we will get our output in YES or NO, we decided to implement models which are for categorical data. To get the most wanted result, we have utilized Logistic Regression, Multinomial Naive Bayes, Random Forest Algorithm , Support Vector Machine (SVM) and LSTM. A brief description of our used methods are given below.

3.5.1 Logistic Regression

Logistic Regression performs classification operations in the algorithm. The equation for logistic regression is:

$$f(x) = \frac{M}{1 + e^{-k(x-x_0)}} \tag{3.1}$$

Here,
e = Euler's number
x$_0 = xvalueofthesigmoid'smidpoint$
$M = maximumvalueofcurve$
$k = steepnessorlogisticrateofthecurve$
$andf(x) = functionoutput.$[1]

LR is used to bunch full scale subordinate data by utilizing pointer factors. It is in like manner called logit method or strategic method, is a by and large used model to look at the association between various free factors and one downright ward variable [15]. It changes over probability scores from full scale subordinate elements, in this way making a relationship between total variable which is dependent and a persevering variable which is free. Binary logistic Regression and multinomial LR are two sorts of LR [1]. The fundamental separation between these two sorts is the sort of imprints used in them.Multinomial determined backslide is performed when the imprints integrate a couple of characteristics. The model's one of the critical assumptions is that the free and subordinate variables don't share an immediate relationship.In our task we have utilized LR since it is easy to comprehend, simple to execute, and proficient to prepare our dataset. It additionally performs well as our dataset is directly distinguishable.

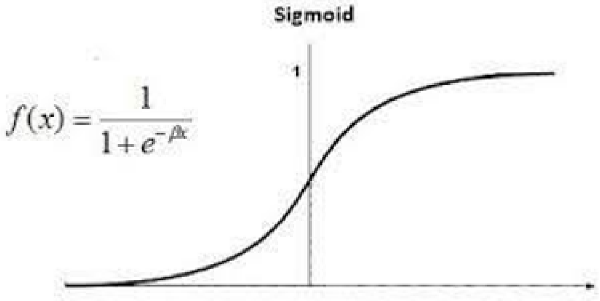


Figure 3.2: logistic regression
[23]

3.5.2 Multinomial Naive Bayes

MNB calculation is a kind of approach that utilizes probabilistic learning. Bayes rule is utilized to sort information by picking the class which most likely have made the occasion [2]. The Bayes hypothesis, finishes up the probability of an occasion happening spread out on past data of the occasion's conditions. It depends upon the situation under. For any components x and y,

$$p(x \mid y)=\frac{p(x) p(y \mid x)}{p(y)}=\frac{p(x, y)}{p(y)} \tag{3.2}$$

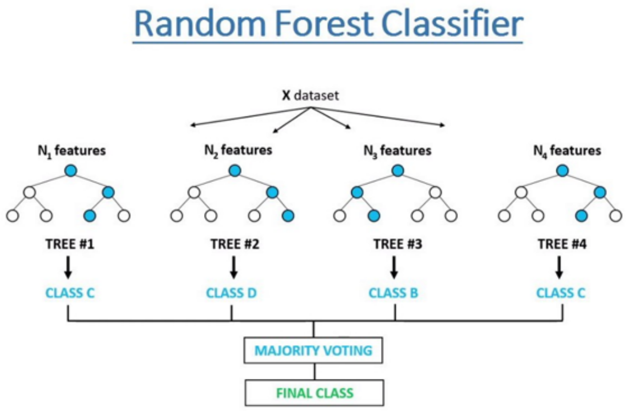
Below it shows Multinomial Naive Bayes model-

$$p(x \mid y)=\frac{p(x) \prod_{i=1}^{n} p\left(w_{i} \mid x\right)^{f i}}{p(d)} \tag{3.3}$$

It is a fundamental approach for building classifiers, models that commit class names to give models, tended to as vectors of part regards, where the class marks are drawn from some confined set [31]. There is certainly not a solitary calculation for arranging such classifiers, yet a get-together of assessments thinking about a conventional rule: all genuine Bayes classifiers expect that the worth of a specific part is freed from the worth of another part, given the class variable [24]. In different pragmatic applications, limit evaluation for MNB models involves the methodology for most absurd probability; in this way, one can work with the Bayes model without getting through Bayesian likelihood or utilizing any Bayesian techniques[31]. Regardless of what their MNB and obviously reshaped questions, clear Bayes classifiers have worked marvelously in different puzzled authentic circumstances. In 2004, an evaluation of the Bayesian social event issue showed that there are sound speculative explanations behind the clearly unrealistic possibility of direct Bayes classifiers [9]. Still, a cautious association with other depiction calculations in 2006 showed that Bayes depiction is outsmarted by different methods, as maintained trees or irregular timberland regions.

3.5.3 Random Forest

Random Forest Algorithm is an outfit learning procedure for arrangement, relapse and different assignments [2]. It was proposed by L. Breiman in 2001 and has been extraordinarily successful as a generally helpful gathering and backslide strategy [8]. In the paper, the makers have proposed a model in predicting future evaluated glomerular filtration rate (eGFR) values, which relies upon Irregular Timberland relapse that can successfully acquire from this current reality EMR data and exactly expect future patient outcomes. The evaluated eGFRs were used to organize patients into CKD stages with high full scale showed up at the midpoint of and smaller than usual tracked down the center worth of estimations. In Random Forest Algorithm examination, the predicting precision was improved by smoothing out the potential gains of hyperparameters. The computational examination achieved an ordinary R*R of 0.95 over three years with little assortment. Besides, a 88 percent Large scale Review and a 96 percent Large scale exactness by averaging more than three years. This is a methodology, which joins a couple of randomized decision trees and sums their conjectures by averaging, has shown extraordinary execution in settings where the quantity of variables is much greater than the amount of discernments [8]. Additionally, it is adequately versatile to be applied to huge extension issues, is really changed in accordance with various unrehearsed learning tasks, and returns extents of variable importance.The primary explanation for involving RF in our work as it lessens overfitting issue in decision tree and furthermore decreases the change and hence works on the exactness.



[33]

Figure 3.3: Random forest

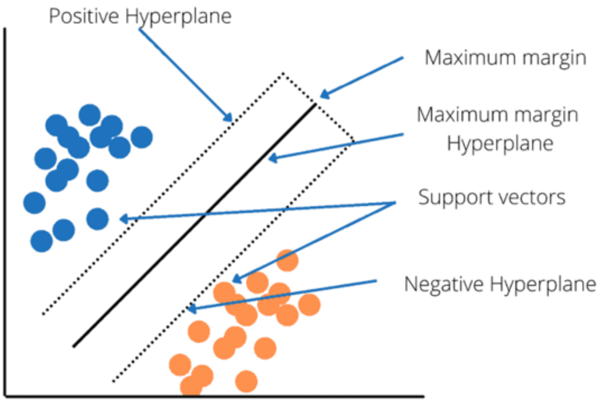
3.5.4 Support Vector Machine (SVM)

SVM is such a model that depends on regulated learning. It's essentially attempting to find a cutoff two-layered space between two unique information focuses record class. This constraint is known as the hyperplane. As a general rule, the design is to find a hyperplane that does this boost the separation of data of interest from related information focuses A class in n-layered space [13]. Next information point

It is known as a help vector to this hyperplane. Since it's paired the characterization hyperplane condition is-

$$w * x - b = 0 \tag{3.4}$$

Here, w is the standard course of the hyper-plane, b is a limit structure and the worth of x can shift for various examples. For a specific information point w, on the off chance that the condition becomes positive, w will have a place with a class. In the event that the condition becomes negative, w will have a place with another class. Typically, multiclass order isn't upheld by SVM. It upholds double order. However, SVM can uphold multiclass order by utilizing two unique methodologies.



[27]

Figure 3.4: SVM model

One-against one and one-versus all.SVM functions admirably when there is a reasonable edge of partition between classes. It is more powerful in high layered spaces. SVM is somewhat memory proficient. Those are reasonable for are work and we picked SVM.

3.5.5 Long Short Term Memory(LSTM)

The name LSTM alludes to the relationship that standard RNNs have both ”long haul memory” and ”transient memory” [29]. The loads and inclinations of the associations inside the organization change once for each preparing set. This might be very similar to how physiological changes in synaptic strength save recalling. The initiation design inside the organization changes once per step. This could be practically similar to how transient changes in release designs inside the mind safeguard recollecting. The LSTM engineering expects to supply present moment or ”long haul memory” for RNNs which will last a huge number of it slow advances [5]. An ordinary LSTM unit comprises of cells, an info door, a result entryway, and a neglect door [30]. The cells store values at inconsistent spans, and furthermore the three doors control the progression of information into and out of the cell. The most important phase in our LSTM is to come to a choice what data we'll eliminate from the cell state [30]. This choice is shaped by a sigmoid layer alluded to as the ”neglected entryway layer”. it's at ht 1 and xt, and results assortment somewhere in the range of 0 and 1 for each number in cell state Ct 1. the sum 1 addresses "keep this totally" while the sum 0 addresses "sort of" drop this by and large".

$$\mathrm{ft} = \sigma(\mathrm{wf}[\mathrm{ht} - 1, \mathrm{xt}]) + \mathrm{bf} \tag{3.5}$$

The subsequent stage is to make your psyche up what new data we will store inside the cell state [30] .This has two sections. Initial, a sigmoid layer called "door layer" concludes what values we will refresh. Then, a tanh layer produces a vector whose new up-and-comer values, C t, might be added to the state. inside the subsequent stage, we'll join these two components to shape a standing update.

$$\mathrm{it}\ \ = \sigma(wi[ht - 1, xt]) + bi \tag{3.6}$$

$$\mathrm{ct} = \tanh(\mathrm{wc}[\mathrm{ht} - 1, \mathrm{xt}]) + \mathrm{bc} \tag{3.7}$$

Presently the time has come to refresh the old cell state to the new Ct cell express The past advances we have previously chosen what to attempt to do, we simply should screw.We increase the old state by pi, failing to remember things we chose to forget before. Then, at that point, we add Ct to that. These are the new applicant values, scaled by the degree to which we consider to refresh each state esteem.

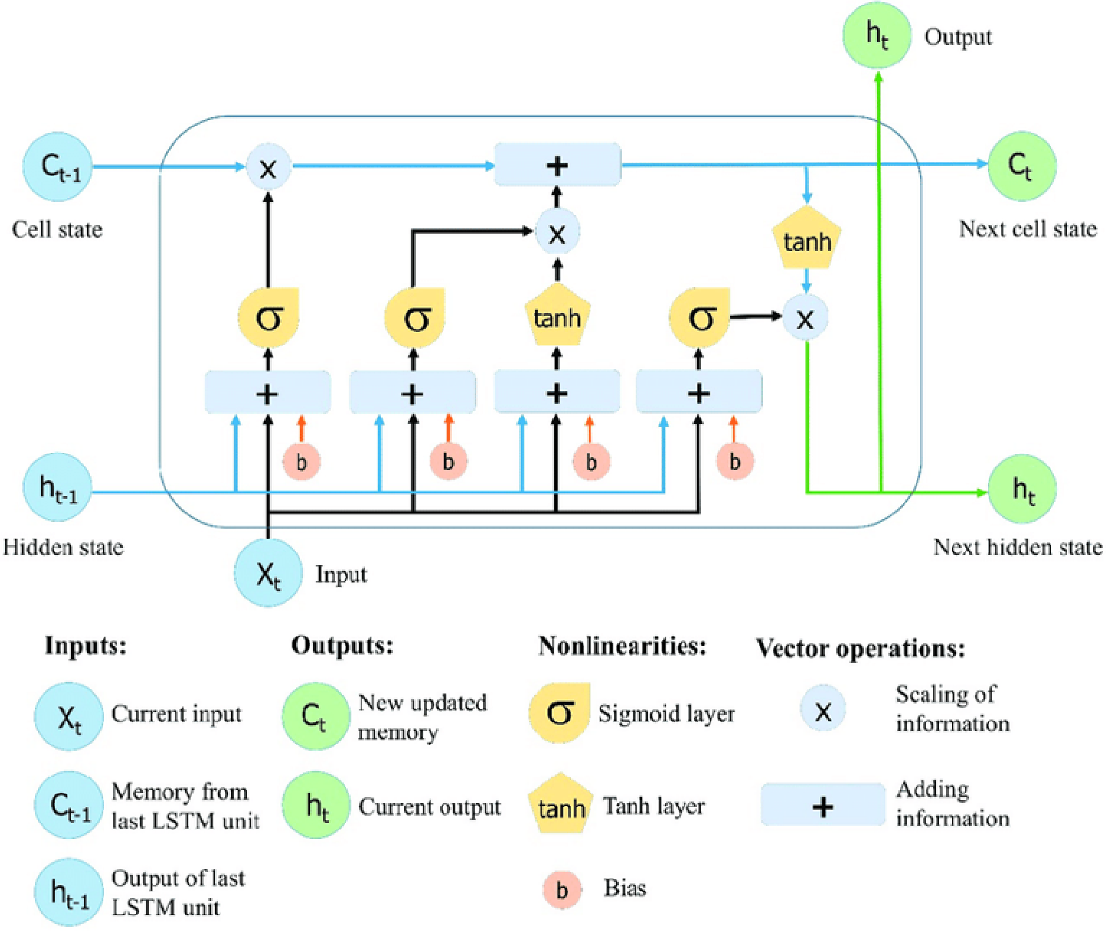
$$\mathrm{ct} = \mathrm{ft} * \mathrm{ct} - 1 + \mathrm{it} * \mathrm{ct} \tag{3.8}$$

Eventually, we needed to make your psyche up the thing we were visiting produce. This result are upheld the condition of our cell, however will be a separated rendition. In the first place, we run a sigmoid layer that concludes which a piece of the cell state we will show. Then, at that point, we set the condition of the cell by means of tanh (to push the qualities from - 1 to 1) and duplicate it by the result of the sigmoid entryway, all together that we create just the parts we chose.

$$\mathrm{ot}\ \ = \sigma(\mathrm{w}[\mathrm{ht} - 1, \mathrm{xt}]) + \mathrm{b} \tag{3.9}$$

$$\mathrm{ht} = \ \ \mathrm{ot}\ \ * \tanh(\mathrm{ct}) \tag{3.10}$$

LSTM networks are viable for grouping, handling and guaging upheld measurement information. Without a doubt, there are in many cases postponements of uncertain time between sequentially critical occasions. LSTM is intended to beat the broken angle issue that happens while preparing conventional RNNs. The relative harshness toward hole length might be an or more of LSTM over RNNs, stowed away Markov models, and other grouping learning techniques in numerous applications.In simple words, the thought is to utilize one LSTM to peruse the info grouping, each timestep in turn, to get huge fixed-layered vector portrayal, and afterward to utilize one more LSTM to remove the result succession from that vector. If we look at all the previous works related to our project, we will see that in those works the use of LSTM did not get such importance. But in our work we want to give more focus on LSTM with other used algorithms. If we get a good output of LSTM in our dataset then it will be an outstanding achievement in the prediction of CKD. We can then also make a hybrid model and use it in a large dataset to get a higher accuracy.



[34]

Figure 3.5: LSTM Model

Chapter 4

4.1 Results

In the wake of fitting the handled dataset into Google Colab, we have got the accuracy for Multinomial Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest Algorithms, LSTM. Among them LSTM has given the most accuracy of 0.95 percent.In the table below we have shown the accuracies of our applied machine learning methods in our training data.After our model has been handled by utilizing the training set which is 80 percent of our total data, we can test the model by making predictions against the testing set.

Implemented models	Accuracy
Support Vector machine	0.94
Random Forest Algorithm	0.95
Logistic Regression	0.95
Multinomial Naive Bayes	0.95
LSTM	0.96

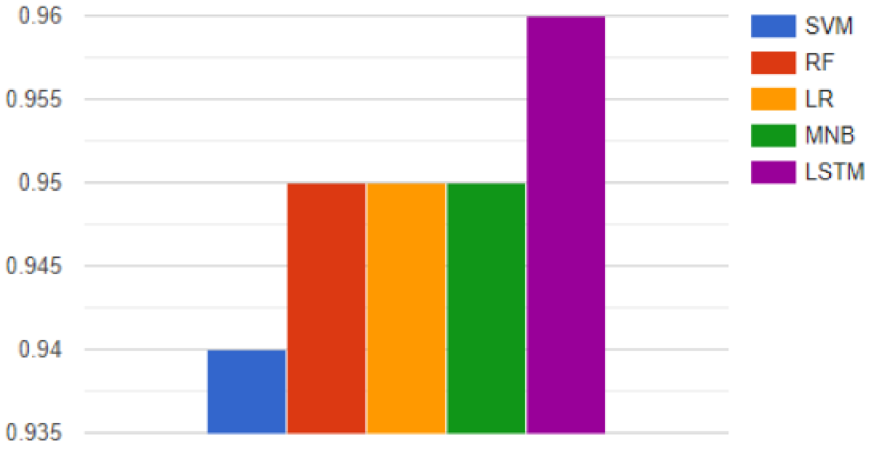


Figure 4.1: Bar chart of training result

After that we went for our testing dataset and got the following accuracies. For this we had to utilize the testing data which is 20 percent of our total data to give an unprejudiced assessment of a last model fit on the training dataset.

Implemented models	Accuracy
Support Vector machine	0.89
Random Forest Algorithm	0.94
Logistic Regression	0.9
Multinomial Naive Bayes	0.91
LSTM	0.95

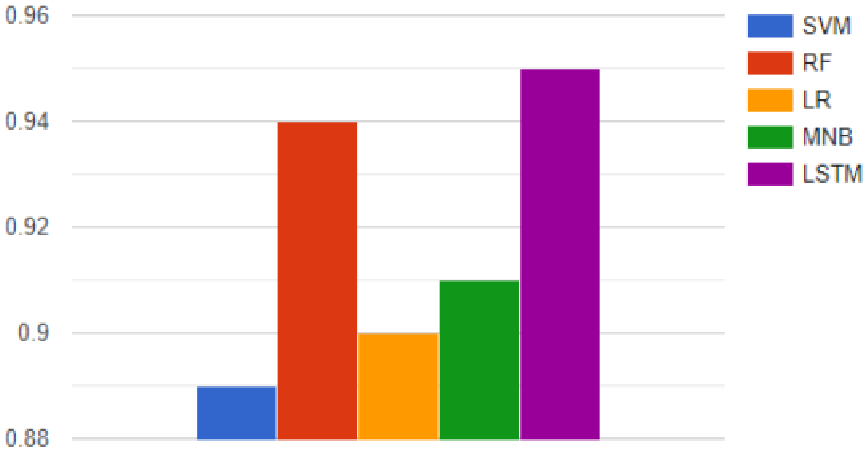


Figure 4.2: Bar chart of testing result

In ML, improving a model characterizes tracking down the best arrangement of hyperparameters for a particular issue. The contrast between model boundaries and model hyperparameters is that, model boundaries are the ones that the model gets the hang of during training, helps in making predictions. On the other hand, model hyperparameters are best-considered settings for a ML algorithm that assists with the learning experience that the information researcher tunes prior to preparing. If we look at our testing results in the testing table and bar chart for presenting testing results we can clearly see that most accuracy that has given is LSTM. To get this highest result from LSTM and overall a good prediction from our other implemented models we had to run our models several times. For our each algorithm we picked the highest accuracy that we got from them while implementing them on our dataset for several times. We have imported all the necessary python libraries for our work at google colab. Below we have shown accuracies and loss graph of our applied algorithms which will help us to understand the result better.

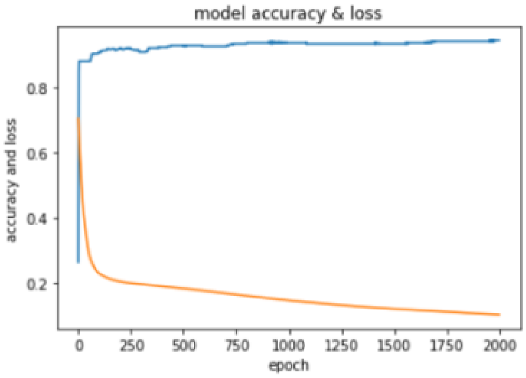


Figure 4.3: Accuracy and loss graph of SVM

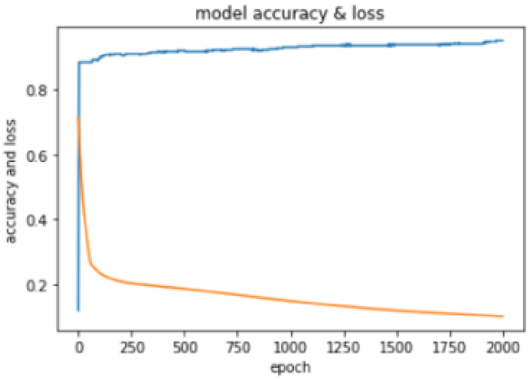


Figure 4.4: Accuracy and loss graph of Random forest

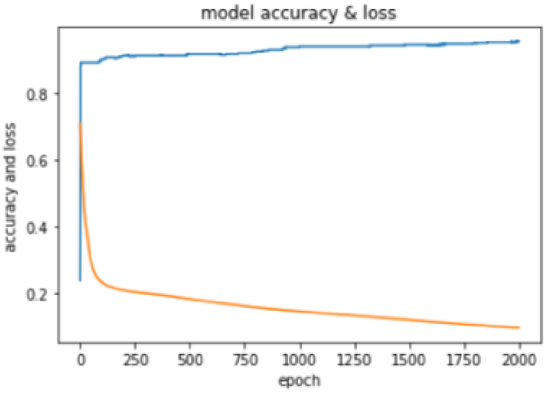


Figure 4.5: Accuracy and loss graph of Logistic regression

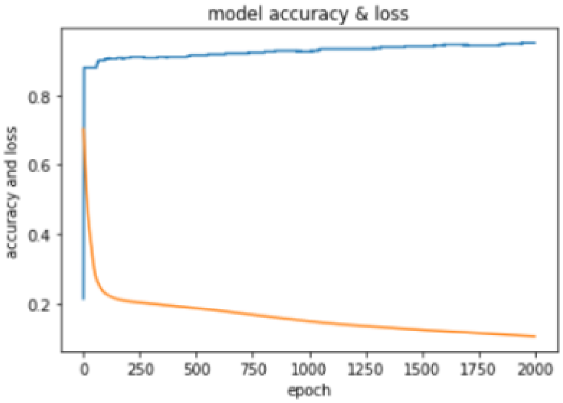


Figure 4.6: Accuracy and loss graph of Naive bayes

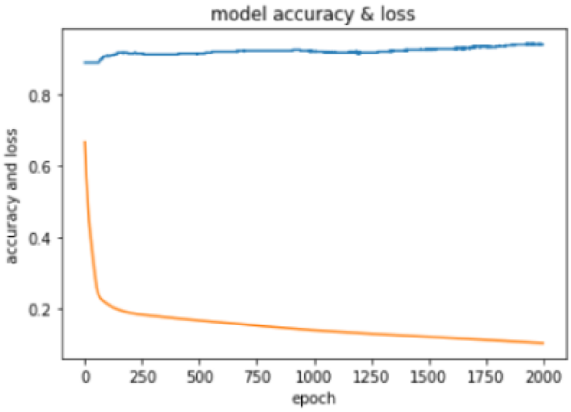


Figure 4.7: Accuracy and loss graph of LSTM

Chapter 5

5.1 Results analysis

Now we want to present our results using confusion matrix which will give us a good understanding of our result. We will use confusion matrix table to describe the performance of our classification model on our set of test data for which the true values are known. If we compare our predicted values with true values of dataset their can be four possible output. One is both of our predicted value and real value is same which is yes-yes, another one is also same but this time both are no-no. And then it can be yes for our predicted value and no for our real value which is yes-no. It also can be no for our predicted value and yes for our real value which is no-yes. We will compare our testing value which is from 20 percent of our dataset which is 99, with real values to show a easy view of our implemented model's accuracy.

First, if we look at the confusion matrix table for SVM we can see that 89 of our predicted values and real values are same. Among them 86 values are no-no. Which means SVM predicted that 86 people do not have CKD and in real dataset exactly those people really did not have CKD. Then, 3 values are yes-yes which means, SVM predicted that 3 people have CKD and in real dataset exactly those people had CKD. Then, SVM predicted 3 people have CKD but in real dataset they have no CKD, which is yes-no. And then, SVM predicted 7 people have no CKD but in real dataset they have CKD, which is no-yes. Now, if we look at the confusion matrix

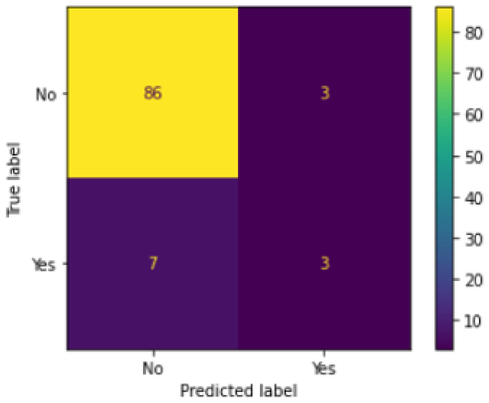


Figure 5.1: Confusion matrix of Support Vector Machine

table for Random Forest Algorithm we can see that 94 of our predicted values and real values are same. Among them 88 values are no-no. Which means RF predicted that 88 people do not have CKD and in real dataset exactly those people really did not have CKD. Then 6 values are yes-yes which means RF predicted that, 6 people have CKD and in real dataset exactly those people had CKD. And then, RF predicted 5 people have no CKD but in real dataset they have CKD which is no-yes. Here, yes-no part is 0 as comparison between 89 values have already completed.

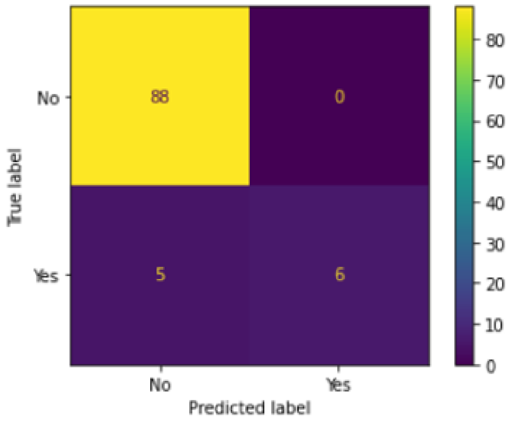


Figure 5.2: Confusion matrix of Random Forest Algorithm

Then if we look at the confusion matrix table for Logistic Regression we can see that 90 of our predicted values and real values are same. Among them 85 values are no-no. Which means LR predicted that, 85 people do not have CKD and in real dataset exactly those people really did not have CKD. Then, 5 values are yes-yes which means LR predicted that 5 people have CKD and in real dataset exactly those people had CKD. And then, LR predicted 9 people have no CKD but in real dataset they have CKD which is no-yes. Here also, yes-no part is 0 as comparison between 89 values have already completed.

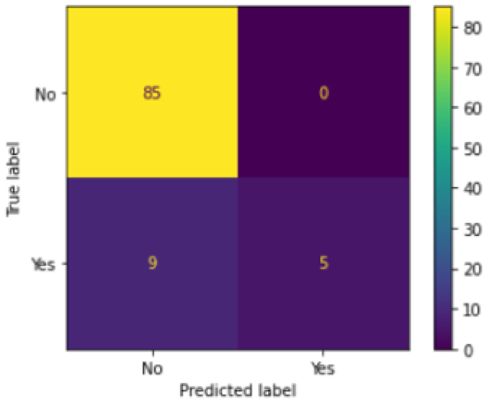


Figure 5.3: Confusion matrix of Logistic Regression

Now, if we look at the confusion matrix table for Multinomial Naive Bayes, we can see that 91 of our predicted values and real values are same. Among them 90 values are no-no. Which means MNB predicted that, 90 people do not have CKD and in real dataset exactly those people really did not have CKD. Then, 1 value is yes-yes which means MNB predicted that 1 person has CKD and in real dataset exactly that person had CKD. And then, MNB predicted 8 people have no CKD but in real dataset they have CKD which is no-yes. Here also, yes-no part is 0 as comparison between 89 values have already done.

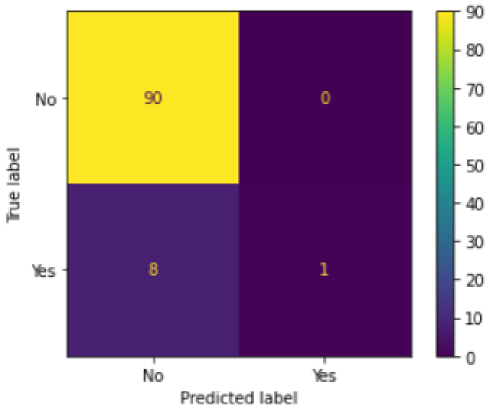


Figure 5.4: Confusion matrix of Multinomial Naive Bayes

Finally, if we look at the confusion matrix table for LSTM we can see that 95 of our predicted values and real values are same. Among them 853 values are no-no. Which means LR predicted that, 83 people do not have CKD and in real dataset exactly those people really did not have CKD. Then, 12 values are yes-yes which means LR predicted that 12 people have CKD and in real dataset exactly those people had CKD. And then, LR predicted 4 people have CKD but in real dataset they have no CKD which is yes-no. But here, no-yes part is 0 as comparison between 89 values have already completed. After all our efforts and challenges we could successfully

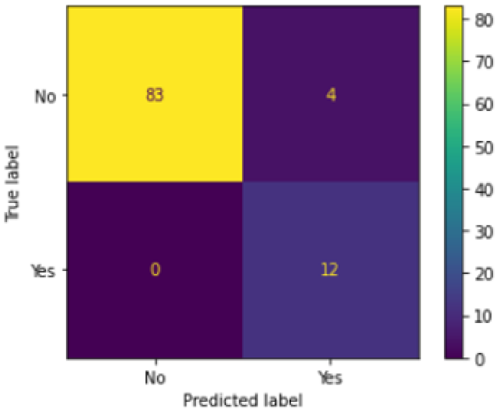


Figure 5.5: Confusion matrix of LSTM

implemented our algorithms. But here the most noticeable and a unique thing is the successful implementation of LSTM. Later we can use it in a larger dataset. We can also make a hybrid model of our task. It will improve our system's performance.

And others can also use this to improve their system. Hopefully it will add a great path in the process of prediction of CKD.

5.2 Paper comparison

After all our efforts and challenges we could successfully implement our algorithms. But here the most noticeable and a unique thing is the successful implementation of LSTM. If we see the works that has been performed in the prediction of CKD a dataset was used which was little old and it had some missing values. In this work, we could find a new real dataset which had no missing values, suitable for in te prediction of CKD. Then some very good works were performed with algorithms like logistic regression, KNN, K-means, random forest, SVM etc. Their accuracy was so good that it went near almost 100 percent. Some good works were also performed with hybrid models. But a noticeable thing is that in all works the implementation of LSTM did not get such priority. With some other algorithms we also wanted to know how effective LSTM is in the prediction of CKD using our new dataset which we have used here. And later we have found that LSTM works very accurately in this field. This is a noticeable achievement for us. There also might be a few potential constraints in this review. In our work, we wanted to find an unique and large dataset. Yes, our dataset in suitable but it could be larger. We were very close to show best accuracies in this context but it could little better. Data visualization could be improved. Later we can use it in a larger dataset. We can also make a hybrid model of our task. It will improve our system's performance. And others can also use this to improve their system. Hopefully it will add a great path in the process of prediction of CKD.

Chapter 6

6.1 Conclusion

This review explores the capability of ML algorithms to distinguish CKD utilizing the least potential tests or elements. Since the information utilized in this study is restricted, we intend to approve our discoveries utilizing a bigger dataset or contrast the outcomes and another dataset that contains similar highlights from here on out. Likewise, to support the decrease of CKD pervasiveness, we desire to utilize important dataset to figure whether an individual with CKD risk factors like as diabetes, hypertension, and a family background of kidney disappointment would foster CKD later on or not.

Bibliography

[1] R. E. Wright, "Logistic regression.," 1995.

[2] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *Australasian Joint Conference on Artificial Intelligence*, Springer, 2004, pp. 488–499.

[3] Q.-L. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: Systematic review," *BMC public health*, vol. 8, no. 1, pp. 1–13, 2008.

[4] R. Lozano, M. Naghavi, K. Foreman, *et al.*, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the global burden of disease study 2010," *The lancet*, vol. 380, no. 9859, pp. 2095–2128, 2012.

[5] D. Monner and J. A. Reggia, "A generalized lstm-like training algorithm for second-order recurrent neural networks," *Neural Networks*, vol. 25, pp. 70–83, 2012.

[6] T. Di Noia, V. C. Ostuni, F. Pesce, *et al.*, "An end stage kidney disease predictor based on an artificial neural networks ensemble," *Expert systems with applications*, vol. 40, no. 11, pp. 4438–4445, 2013.

[7] J. Radhakrishnan, G. Remuzzi, R. Saran, *et al.*, "Taming the chronic kidney disease epidemic: A global view of surveillance efforts," *Kidney international*, vol. 86, no. 2, pp. 246–250, 2014.

[8] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[9] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 management and innovation technology international conference (MITicon)*, IEEE, 2016, MIT–80.

[10] K. A. Padmanaban and G. Parthiban, "Applying machine learning techniques for predicting the risk of chronic kidney disease," *Indian Journal of Science and Technology*, vol. 9, no. 29, pp. 1–6, 2016.

[11] R. Ruiz-Arenas, R. Sierra-Amor, D. Seccombe, *et al.*, "A summary of worldwide national activities in chronic kidney disease (ckd) testing," *Ejifcc*, vol. 28, no. 4, p. 302, 2017.

[12] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, *et al.*, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *2018 IEEE congress on evolutionary computation (CEC)*, IEEE, 2018, pp. 1–9.

[13] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (svm) learning in cancer genomics," *Cancer genomics & proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[14] P. Scholar, "Chronic kidney disease prediction using machine learning," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 16, no. 4, 2018.

[15] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *International Journal of Soft Computing and Its Applications*, vol. 10, no. 8, 2019.

[16] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20 991–21 002, 2019.

[17] S. Revathy, B. Bharathi, P. Jeyanthi, and M. Ramesh, "Chronic kidney disease prediction using machine learning models," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 6364–6367, 2019.

[18] R. S. Sealfon, L. H. Mariani, M. Kretzler, and O. G. Troyanskaya, "Machine learning, the kidney, and genotype–phenotype analysis," *Kidney international*, vol. 97, no. 6, pp. 1141–1149, 2020.

[19] D. Chicco, C. A. Lovejoy, and L. Oneto, "A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease," *IEEE Access*, vol. 9, pp. 165 132–165 144, 2021.

[20] M. M. Nishat, F. Faisal, R. R. Dip, *et al.*, "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 29, e1–e1, 2021.

[21] M. Ahmed, M. Ali, N. Ahmed, T. Bhuiyan, *et al.*, "Computational intelligence approaches for prediction of chronic kidney disease," in *Advances in Distributed Computing and Machine Learning*, Springer, 2022, pp. 299–309.

[22] *Acute kidney injury versus chronic kidney disease — nursingcenter*, https://www.nursingcenter.com/ncblog/january-2020/acute-kidney-injury-and-chronic-kidney-disease, (Accessed on 09/20/2022).

[23] *Advantages and disadvantages of logistic regression - geeksforgeeks*, https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/, (Accessed on 09/25/2022).

[24] *Chronic kidney disease ehrs abu dhabi — kaggle*, https://www.kaggle.com/datasets/davidechicco/chronic-kidney-disease-ehrs-abu-dhabi, (Accessed on 09/20/2022).

[25] *Chronic kidney disease: The silent killer*, https://www.sarojhospital.com/blog/chronic-kidney-disease-the-silent-killer, (Accessed on 09/19/2022).

[26] *Global facts: About kidney disease — national kidney foundation*, https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease, (Accessed on 09/20/2022).

[27] *Implementation of support vector machine (svm) using python*, https://hands-on.cloud/implementation-of-support-vector-machine-svm-using-python/, (Accessed on 09/25/2022).

[28] *Kidney scan: Purpose, procedure, risks, results*, https://www.webmd.com/a-to-z-guides/kidney-scan-what-to-expect, (Accessed on 09/20/2022).

[29] *Lstm — introduction to lstm — long short term memor*, https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/, (Accessed on 09/20/2022).

[30] *Lstm — introduction to lstm — long short term memor*, https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/, (Accessed on 09/20/2022).

[31] *Multinomial naive bayes classifier algorithm*, https://www.mygreatlearning.com/blog/multinomial-naive-bayes-explained/, (Accessed on 09/20/2022).

[32] *Patient education: Dialysis or kidney transplantation — which is right for me? (beyond the basics) - uptodate*, https://www.uptodate.com/contents/dialysis-or-kidney-transplantation-which-is-right-for-me-beyond-the-basics, (Accessed on 09/20/2022).

[33] *Random forest classifier and its hyperparameters — by ankit chauhan — analytics vidhya — medium*, https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6, (Accessed on 09/25/2022).

[34] *The structure of the long short-term memory (lstm) neural network.... — download scientific diagram*, https://www.researchgate.net/figure/The-structure-of-the-Long-Short-Term-Memory-LSTM-neural-network-Reproduced-from-Yan_fig8_334268507, (Accessed on 09/20/2022).