Sepsis Diagnostic System

*Submitted in partial fulfilment of the requirements*

*of the degree of*

BACHELOR OF ENGINEERING

*in*

INFORMATION TECHNOLGY

(A.Y. 2018-2019)

by

|  |
| --- |
| Mukesh Choudhary (Roll No.:13)  Roshan Choudhary (Roll No.:14)  Chinmay Dali (Roll No.:15) |

Under the Guidance of

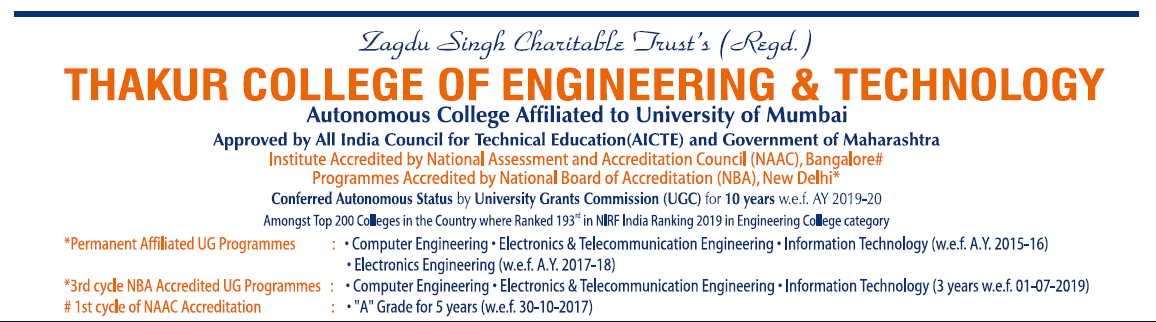
Mrs. Sudhir Dhekane

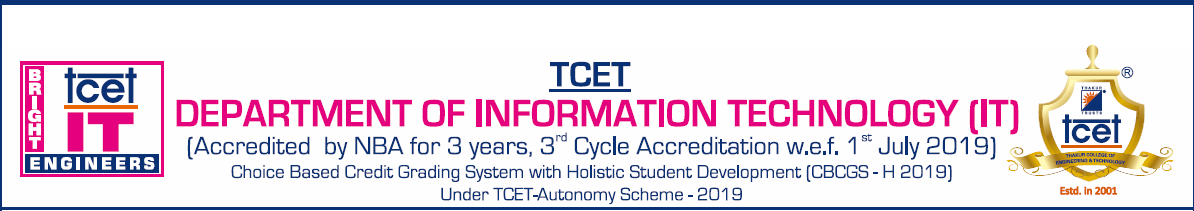
Assistant Professor, I.T Department, TCET

****

**Choice Based Credit Grading System with Holistic Student Development**

**(CBCGS-H 2019)**

****

****

**Certificate**

This is to certify that Mr./Ms.\_\_\_\_\_\_\_\_\_, Mr./Ms.\_\_\_\_\_\_\_, Mr./Ms.\_\_\_\_\_\_\_\_\_,

are bonafide students of Information Technology Department, Thakur College of Engineering and Technology, Mumbai. They have satisfactorily completed the requirements of PROJECT-I as prescribed by **Thakur College of Engineering and Technology (An Autonomous College affiliated to University of Mumbai),** while working on “Project Title”.

|  |  |
| --- | --- |
| Signature :------------------------------Name : Sudhir Dhekane  Designation | Signature :------------------------------  Name : Dr. Rajesh S. Bansode  HOD-IT |
| Signature: ---------------------------------------------  Name : Dr. B. K. Mishra  Principal,  Thakur College of Engineering and Technology. | |

|  |
| --- |
| **Internal Examiner: External Examiner:** |
| Signature :--------------------------------- Signature :---------------------------------  Name : Name : |
| Thakur College of Engineering and Technology, Kandivali (East) Mumbai.  Date:  Place: |

**DECLARATION**

I/we declare that this written submission represents my/our ideas in my/our own words and where others ideas or words have been included, I/we have adequately cited and referenced the original sources. I/we also declare that I/we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my/our submission. I/we understand that any violation of the above will because for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1. -----------------------------------------

Mukesh Choudhary (Roll No.: 13)

1. ---------------------------------------

Roshan Choudhary (Roll No.: 14)

1. -----------------------------------------

Chinmay Dali (Roll No.: 15)

Date:

**ACKNOWLEDGEMENT**

We sincerely thank to our guide Mrs. Pranjali Kasturi for her guidance and support for carrying out our project work. We also thank our guide for giving us new ideas to develop our project well. We also give our sincere thanks to our Principal Dr. B. K. Mishra, Vice-Principal Dr. Deven Shah, HOD-IT Dr. Rajesh Bansode, our project coordinators for arranging the necessary facilities to carry out the project work.

We thank Thakur College of Engineering and Technology for their support and providing us a platform to explore new trending technologies in the world.

1. Mukesh Choudhary (Roll No: 13)
2. Roshan Choudhary (Roll No: 14)
3. Chinmay Dali (Roll No: 15)

**ABSTRACT**

Sepsis is a major cause of death in the world. World Health Organization estimates 30 million people developing sepsis and 6 million people die from sepsis each year; of severe sepsis and ultimately earlier intervention. However, current methods for identifying and predicting severe sepsis are biased and inadequate. The goal of this work is to identify a new framework for the prediction of severe sepsis and identify early predictors utilizing clinical laboratory values and vital signs collected in patients. Early prediction of sepsis is critical for improving sepsis outcomes. The late prediction of sepsis in non-sepsis patients is a challenging problem. The aim of this study is to develop an artificial intelligence-based early warning and therapeutic decision support system which reduces sepsis-associated hospital mortality.

[Commitments, deliverance, result and conclusion ]

|  |  |  |
| --- | --- | --- |
|  | **List of Figures** |  |
|  |  |  |
|  |  | Page No. |
| Fig.2.5.1 | KDD Process | 15 |
| Fig.2.5.2 | Scrum | 17 |
| Fig.3.3.1 | Phases of project development | 21 |
| Fig.3.4.1 | Gnatt Chart | 24 |
| Fig.4.2.1 | Block Diagram | 27 |
| Fig.4.3.1 | Flowchart | 28 |
| Fig.4.4.1 | Database Screenshot | 29 |

**C O N T E N T S**

* Certificate i
* Declaration ii
* Acknowledgement iii
* Abstract iv
* List of Figures v
* List of Tables vi

**Chapter No. Topic Pg. .No.**

**Chapter1 Overview 1**

1.1 Introduction 02

1.2 Background 03

1.3 Importance of the Project 03

1.4 Objectives and Scope of the Project 03

1.5 Summary 04

**Chapter2 Literature Survey & Proposed Work 5**

2.1 Introduction 06

2.2 Literature Survey Table 06

2.3 Problem definition 14

2.4 Features of the Project 14

2.5 Methodology Used 15

**Chapter3 Analysis and Planning 19**

3.1 Introduction 20

3.2 Feasibility Study 20

3.3 Project planning 21

3.4 Scheduling 24

**Chapter4 Design and Implementation 26**

4.1 Model Requirements 27

4.2 Block Diagram 27

4.3 Flow Chart 28

4.4 Database screenshot 29

**Chapter 5 Results & Discussion 30**

5.1 Results 31

5.2 Future Scope 31

**Chapter 6 Conclusion 32**

6.1 Conclusion 33

**References 34**

**CHAPTER 1**

**OVERVIEW**

* 1. **Introduction**

Severe sepsis, a systemic response to infection complicated by organ dysfunction, is a common cause of hospital morbidity and mortality. Much work is being done to

further model and predict the outcomes for patients with signs of infection or those who may be vulnerable to such events. Unfortunately, the systemic inflammatory response

syndrome (SIRS) criteria which are used to diagnose severe sepsis may miss as many as 1 in 8 cases and predict outcomes poorly. Therefore, improved characterization of the systemic inflammatory response and factors that identify and risk-stratify patients with severe sepsis while they are in the hospital are needed. Lactate is the end product of cytosolic glucose metabolism without the use of the mitochondria, a largely

inefficient process. While some tissues function using anaerobic metabolism (such as skeletal muscle), most clinical scenarios in which lactate is overproduced signifies

cellular disarray. Typically, this is due to global or regional hypoxia, or deficits in lactate clearance. Septic shock may occur when oxygen utilization is compromised for a variety

of reasons and lead to hyperlactatemia. Regardless of the physiology, the data are clear that admission lactate concentrations, maximum lactate concentrations, and time to

normalization of lactate concentrations correspond with hospital mortality from severe sepsis. In this work, we define severe sepsis as the presence of 1) suspected infection and 2) hyperlactatemia as evidenced by blood culture acquisition and a concurrent lactate

concentration of at least 4 mmol/L. Using the clinical from UCSD and BIDMC dataset we would be creating an interactive website and mobile app leveraging Expert-System for prediction of medical diagnostics.

* 1. **Background**

The focus concerns, more general development of clinical decision support systems (CDSSs) that, based on the collected patient data from physical exams, laboratory results etc, can provide early warnings to the healthcare professionals Many of the CDSSs used today stem from research on expert systems, where the aim was to build software that could simulate human thinking through mapping signs, symptoms and laboratory results into probabilistic estimates of different diagnoses. The trend nowadays is rather to develop support systems that assist the clinician in his/her decision making, and where the user is active in the decision-making process. These systems range from being purely rule-based to using techniques such as artificial neural networks, support vector machines and decision trees to create models suitable for sepsis diagnosis.

card and „card holder not present‟ fraud.Stolen/lost card fraud occurs when fraudsters steala credit

card and „card holder not present‟ fraud.Stolen/lost card fraud occurs when fraudsters steala credit

* 1. **Importance of Project**

Several studies have concluded that clinical decision support can improve clinician performance. Positive effects concern, for example, the stronger adherence to relevant guidelines, the cost of care, the reduction of medication errors and decreased rates of potential redundant or inappropriate care. However, a growing pool of research indicates that the anticipated positive effects are often unrealized, as well as that the impact

of introducing the CDSS on the workflow of the clinicians has not been sufficiently evaluated. Early clinical recognition of sepsis can be challenging. With the advancement of machine learning, promising real-time models to predict sepsis have emerged. positive

results of introducing a new CDSS are most often found in cases where there

has been an incremental and local development during several years, led by

researchers within the field, and where other computerized support systems are

an accepted part of the work environment.

* 1. **Objectives and Scope of the Project**

The **objectives** of **Medical Diagnostic System** of this paper is to predict sepsis in advance of an episode, thereby creating an expert system for providing the suggestions at a time and location where the decisions are being made provide actionable recommendations and computerize the entire process. This system aids clinicians adhere to regulated treatment plans for sepsis as well as to monitor the patient’s status through visual means. The system fetches real-time patient data from the patient’s records, and the clinician is able to order medications and procedures through the tool.

Scope:

Medical expert systems were initially developed for academic areas and later for clinical applications also. Health care systems produce tremendous amounts of information (patient, demographic, clinical and billing data), which are susceptible to analysis by intelligent software and need new techniques to extract new knowledge. A variety of medical expert systems tools are available and can function as intelligent assistants to clinicians, helping in diagnostic processes, laboratory analysis, treatment protocol, and teaching of medical students and residents. Although to implement one in a production environment we need to follow the guidelines of hospital which is not the scope of the project, despite these differences, and not being the scope of this article, it would be very interesting to test the expert system of sepsis in a hospital environment.

* 1. **Summary**

The chapter consists of the introduction to the concept the background of the title selected and discuss about the why this project is important the scope of it etc.

**CHAPTER 2**

**LITERATURE SURVEY**

**& PROPOSED WORK**

**2.1 Introduction**

A writing literature review surveys overview insightful articles, books, papers, gathering procedures and different assets which are pertinent to a specific issue, zone of research, or hypothesis and gives a setting to an exposition by recognizing past research. Research recounts a story and the current writing causes us to recognize where we are in the story as of now. It is up to that composition a paper to proceed with that story with new research and new points of view yet they should initially be comfortable with the story before they can push ahead.

**2.2 Literature Survey Table**

Table no.2.2.1: Literature Survey Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ref No.** | **Year** | **Authors** | **Title** | **Methodology** | **Key findings** | **Research gaps** |
| [1] | 2015 | Joseph Guillén, Jiankun Liu, Margaret Furr, Tianyao Wang, Stephen Strong, Christopher C. Moore, Abigail Flower, Laura E. Barnes | Predictive Models for Severe Sepsis in Adult ICU Patients | Patient selection  Data cleaing  Database  The square weight  assigned by an SVM  classifier  Ten-fold cross validation  was applied to calculate  he average metric scores.  SVM  Feature set  reduction  Akaike Information  Criterion (AIC) with  backward elimination to  reduce features  Logistic model  trees  Model  Performance  criteria  performs a greedy  search algorithm  author used best-  first search  utilized Bayesian Information  Criterion (BIC) using the  backward elimination for model  selection.  Logistic  regression | significant promise for the development of a new framework for the detection of severe sepsis. SVM outperforms LR and LTM in terms of predictive power. However, SVM lacks the clinical transparency and interpretability of regression and tree based methods. | The author set out to investigate which physiological patterns or clinical events precede severe sepsis but didn’t considered the outcome such as mortality or lactate clearance |
| [2] | 2018 | Md.Mohaimenul Islam,  Tahmina Nasrin,  Bruno Andreas Walther,  Chieh-Chen Wu,  Hsuan-Chia Yang,  Yu-Chuan Li(Jack) | Prediction of sepsis patients using machine  learning approach: A meta-analysis | The primary analysis was divided into two parts:  a) Sepsis prediction 3 to 4 h prior to onset  b) Sepsis detection (0 h)  1. The author tried to generalize the accuracy using I2 value which was used to assess the statistical heterogeneity which provided an estimate of the percentage of variability among the included studies  2. Likelihood ratios were assessed to express how much more frequent the respective result is among the studies with sepsis disease than among subjects without sepsis disease.  3. diagnostic odd ratio (DOR) was calculate to provide, how much greater the odds of having the sepsis disease are for the people with a positive test result than for the people with a negative test result.  4. The confidence intervals of overall sensitivity and specificity was also analyzed using the F distribution method to compute the exact confidence limits for the binomial proportion [normal approximation to binomial] | the diagnosis of sepsis patients is always challenging due to preexisting organ dysfunction, treatment prior to admission, and concurrent organ support; therefore patients medical history should be considered too. | Despite many attempts to identify sepsis patients, it is still  difficult for healthcare providers to correctly recognize and diagnose this condition because of the heterogeneous nature of possible infection. The meta-analysis couldn't able to justify which ML model performed better predictions even for the same attributes |
| [3] | 2019 | R Murat Demirer | 1    Early Prediction of Sepsis from Clinical Data Using Artificial Intelligence | 1. design a utility function, 𝑈𝑗 (𝑗 = 1,2) that rewards early predictions and penalizes late predictions as well as false alarms from sequences of different states to real space  2. HEAVY MATHEMATICS \_\_ HOPING SIR WOULD HELP US | Different datasets used in literatureare theoretically formed a new mathematical theory. Author applied the early detection of sepsis to optimal sequential decision-making theory to achieve Pareto Optimality conditions under Sepsis/Non-sepsis beliefs with differing positive and negative utility predictions at a given time 𝑡 (suspicious, optimal and late). | NULL |
| [4] | 2019 | S.P Maniraj | Christopher Bartona, Uli Chettipallya,b, Yifan Zhouc,d, Zirui Jiangc,e, Anna Lynn-Palevskyc, Sidney Lec, Jacob Calvertc, Ritankar Dasc, | 1. created the trees in the Python XGBoost package. Each branch of a tree was split by creating two groups of a feature. To limit branching to four, three and six branches, for 0-h,24-h,and 48h prediction, respectively.  a. They did so on the basis of ﬁve-fold cross-validation grid search on the training set  2. The dataset were split into two groups, 80% of encounters were allocated uniformly-at-random to a training set, and 20% of encounters were randomly allocated to an independent  a. To perform the hyperparameter grid search, they conducted ﬁve-fold cross-validation using only the encounters allocated to the training set  3. They further divided the 80% training set into ten, equally-sized groups for ten-fold cross validation, again allocating the encounters uniformly-at-random  4. feature importance scores generated by XGBoost, which represent the number of times each feature was used to split the data across the trees; these scores were averaged across the ﬁve folds.  5. The cross-population validation experiments were used to evaluate the algorithm's sepsis detection performance | Other studies have applied machine learning techniques to predict sepsis onset up to 48h prior to organ failure, although these algorithms require signiﬁcantly more data input including patient histories and laboratory test results. This study demonstrates that high sensitivity and speciﬁcity for sepsis can be attained up to 48h prior to organ failure using only commonly measured vital signs, without requiring laboratory tests which may only be ordered after the initial clinical suspicion of sepsis and also abiding ICD codes | The study population excluded patients who presented with sepsis on admission; the algorithm may perform diﬀerently on patients who are admitted with sepsis;  Because we did not require the presence of a sepsis related ICD code for a positive classiﬁcation, it is possible that we positively identiﬁed patients with similar acute vital sign trends but without sepsis. Algorithm performance maybe lower for such patients than for the general population thereby opening the question that those patients had sepsis overtime. The Aftermath of which requires a lot of medical history and research |
| [5] | 2018 | ASWATHY M S, LIJI SAMEUL | Credit Card Fraud Detection using Various Methods and Techniques | In this study mainly two approaches namely misuses (supervised) and anomaly detection (unsupervised) technique is being used. After this a classification is also used for checking the capability to process categorical and numerical data. In the first approach the data is classified as fraud based on previous data. With the help of this dataset classification models are also created, which can predict whether the data is fraud or not. | Hidden Markov Model  and SOM (Self Organizing Map) is used. | Neural networks are the better way to predict and analyze things in a better way and in a more accurate way with the help of various parameters. |
| [6] | 2019 | Joonghee Kim a,1 , HyungLan Chang b,1 , Doyun Kim a , Dong-Hyun Jang a , Inwon Park a , Kyuseok Kim c | Credit Card Fraud Detection | we assumed four scenarios of different predictor  availability: 1) baseline predictors (age, sex, vital signs, SpO2 and  AVPU), 2) baseline predictors and CC embedding, 3) baseline predictors  plus initial laboratory test results, and 4) baseline predictors plus both  (CC embeddings and laboratory test results).  All ML algorithms were  optimized and tested in these four different conditions.  The study population was randomly partitioned into training and  test datasets at a ratio of 6:4 using patient identifification numbers  Six  base ML algorithms, including support vector machine with radial  basis function kernel (SVM), gradient-boosting machine with  Bernoulli loss (GBM), random forest (RF), multivariate adaptive regres  sion splines (MARS), least absolute shrinkage and selection operator  (lasso) and ridge regression, were assessed. In addition, we also con  structed two ensemble classififiers utilizing SVM, GBM, RF, MARS and  lasso as base learners (ridge was not utilized given its high correlation  with lasso). | We observed these clas  sififiers have high discriminatory power even when provided with only  baseline data and outperform traditional scores, such as qSOFA or  MEWS.  MLP classififiers consis  tently required only one or two layers for best performance in various  conditions of predictor availability.  grid search is an ineffificient method to optimize  the hyperparameters and consumes considerable computing resource  and time.  The performance of ML classififiers  was high enough for practical use. Ensembles of base classififiers showed  the best performance and additional information from CC embedding  provided relatively small gain. | it consistently  underperformed compared with tree-based classififiers (e.g., GBM and  RF) and was not improved by inclusion of CC embedding.  the study population and the outcome events (septic shock) were based on time  stamped EHR records instead of clinician's comprehensive evaluation.  patients' under  lying conditions were not incorporated into predictors. |
| [7] | 2019 | R Murat Demirer  Oya Demirer | Early Prediction of Sepsis from Clinical Data Using  Artificial Intelligence | We proposed Partially  Observed Markov Decision Process (POMDP) based on what  we challenge to predict sepsis approximately 6 hours,  before the clinical prediction of sepsis in which measurements.  We design a utility function, () that  rewards early predictions and penalizes late predictions as well  as false alarms from sequences of different states to real space,  ℝ. We automatically aim to identify a patient's risk of sepsis  and make a positive or negative prediction of sepsis for every  time interval. We make positive or negative predictions of  utility functions ((, () for sepsis or as utility functions ((, () for non-sepsis  patient for every time instant during the progression of states.  denotes the patient. We give positive and negative  predictions for obtaining the highest utility score for the patients.  The probabilistic distribution of spectral values with  different beliefs (Sepsis/Non-sepsis) are obtained from the  Lomb-Scargle periodogram. Our spectral model is based on POMDS  Markov process in which patient states are controlled by  treatment on the state transitions that cannot be measured  exactly.  Early detection and antibiotic treatment of sepsis are  modeled by POMDS model for improving sepsis outcomes in  times to treatment represented as hidden states 1, 2, 3, 4 ,  where each hour of delayed treatment has been associated with  roughly an 4-8% increase in mortality [4]. 1 denotes clinical  suspicion of infection at a time, after admission to  hospital and 4 denotes the state of patient at a late time. | Methods such as deep  learning cannot be enough to identify a patient's risk of sepsis  and make a positive or negative prediction of sepsis for every  time interval  Deep neural network models that have been implemented  so far can capture common patterns acquired from thousands of  patients in a supervised training approach. Then, this  supervised neural network approach ignores personalized  sepsis dynamics and requires huge training sets apriori. | Sepsis and non-sepsis beliefs increase  the degree on which the machine’s decisions will affect the  utilities. In addition, the onset time of sepsis will be learned by  the POMDP rather than explicitly specified. |
| [8] | 2019 | James Morrill  , Andrey Kormilitzin  , Alejo Nevado-Holgado  , Sumanth Swaminathan  , Sam  Howison  , Terry Lyons | The Signature-Based Model for Early Detection of Sepsis  From Electronic Health Records in the Intensive Care Unit. | The method takes in data se  quentially and uses the signature transformation to turn the  time-series data into useful features. These features are  fed, along with the variable information at the current time  point, into a gradient boosting algorithm to learn combina  tions of features relevant to sepsis, which then leads to a  risk score for the patient. | We showed that the signature rep  resentation produced a useful summary of the longitudinal  physiological measurements that was used to effectively  discriminate septic from non-septic cases.  The addition of  the signature terms improved significantly the predictive  algorithm  We saw that patients could be labelled as sepsis risk patients with an AUC of 0.868 score considering the score  against those who do eventually develop sepsis. | the model can be  used as an effective screening tool for sepsis, it does not in  general predict cases in the desired 6-hour window before  they occur |
| [9] | 2016 | Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R. | Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data | We apply InSight, a machine learning classification system that uses multivariable combinations of easily obtained patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age), to predict sepsis using the retrospective Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III dataset, restricted to intensive care unit (ICU) patients aged 15 years or more. Following the Sepsis-3 definitions of the sepsis syndrome, we compare the classification performance of InSight versus quick sequential organ failure assessment (qSOFA), modified early warning score (MEWS), systemic inflammatory response syndrome (SIRS), simplified acute physiology score (SAPS) II, and sequential organ failure assessment (SOFA) to determine whether or not patients will become septic at a fixed period of time before onset. We also test the robustness of the InSight system to random deletion of individual input observations.   |  |  |  | | --- | --- | --- | | **Parameter** | **Sepsis (n=377)** | **Non-sepsis (n=377)** | | HR (bpm) | 84.86 ± 0.12 | 79.25 ± 0.11 | | RR (breaths per minute) | 17.94 ± 0.03 | 17.56 ± 0.02 | | SBP (mmHg) | 128.22 ± 0.17 | 135.79 ± 0.18 | | DBP (mmHg) | 70.55 ± 0.11 | 74.19 ± 0.13 | | Temperature (°C) | 36.87 ± 0.01 | 36.80 ± 0.01 | | Peripheral oxygen saturation (SpO2) | 97.17 ± 0.02 | 96.66 ± 0.04 | | WBC (109/L) | 11.21 ± 0.04 | 9.04 ± 0.06 | | LOS (days) | 11.72 ± 0.89 | 6.64 ± 0.57 | | value-dependent SAPS II and SOFA scores when computed at ICU admission. Using the retrospective MIMIC-III dataset and the new Sepsis-3 definition of sepsis, we trained this system to predict sepsis onset and tested its performance | It is not designed to “discover” a set of rules that could create a manual scoring system. InSight is designed as an automatic, EHR-integrated system. Due to its several sequential calculations, including mapping of the input data to a higher-dimensional feature space, InSight scores are infeasible to calculate by hand. |
| [10] | 2018 | Franco van Wyk, Anahita Khojandi, Akram Mohammed, Edmon Begoli, Robert L. Davis, Rishikesan Kamaleswaran | A Minimal Set of Physiomarkers in High Frequency Real-Time Physiological Data Streams Predict Adult Sepsis Onset Earlier | A data-set consisting of high-frequency physiological data from 1,161 critically ill patients was analyzed. 377 patients had developed sepsis, and had data at least 3 hours prior to the onset of sepsis. A random forest classifier was trained to discriminate between sepsis and non-sepsis patients in real-time using a total of 132 features extracted from a moving time-window. The model was trained on 80% of the patients and was tested on the remaining 20% of the patients, for two observational periods of lengths 3 and 6 hours prior to onset. | Sepsis, Physiological Data, Artificial Intelligence, Predictive Model, Critical Care  . F1 score, accuracy, PPV, sensitivity, specificity, and AUC of the trained RF classifiers in discriminating sepsis and non-sepsis patients when used in an online fashion on separate test sets for 3- and 6-hour observational periods | we also combined these extracted features with those obtained from shorter time intervals, i.e., the last 30 and 15 minutes within the time-window, to emphasize the latest physiological changes |
| [11] | 2017 | Futoma, J. E | Machine Learning Algorithmic and System Level Considerations for Early Prediction of Sepsis | In this work, we have taken a pragmatic approach on how to add value to the sepsis prediction problem.  Detecting sepsis in developing countries implies working in a not-so-data-rich environment; both in terms of features and temporal sampling. The prediction model needs to be minimally complex, so that it is easily tune and adapted to local needs. We set out to find a solution which is incrementally better than the current situation in developing countries, rather than a 100% solution. | We have noted that the hourly-sampled feature data varies widely across patients. Sequential features: Given the clinical importance of vitals and that vital samples are available over 90% of the inputs, we use vital values of HR, O2Sat, SBP, DBP, MAP, Temp, Resp as our primary features. We use min- max normalization for each of the features. | we realize that our model can be improved and we intend to continue refining it by improved how we selected the features, handled missing and noisy features, by defining a more sophisticated risk-factor synthetic feature and by training on more datasets. |

**2.3 Problem Definition**

The main objective of this project is early prediction of sepsis using machine learning, we will be using machine learning and Artificial Intelligence. To enable early diagnosis of sepsis, research from three fronts are being conducted. One is the development of multimarket panels. Data mining techniques are being utilized for the selection of several subsets of combinations of biological markers and clinical data, aiding earlier diagnosis, since the biological markers are monitored directly from the patient’s blood.

to an Expert-System that will be able to --------------

**Phase1:** Data collection and Data cleaning and preprocessing it in a single format so that it can be processed easily without any error. The data mining will provide a classification score for the probability of sepsis as an outcome.

**Phase2:** the second main effort focuses on methods for explaining and visualizing results from the data mining process, aiming to increase the interpretability and transparency of the sepsis diagnosis support system.

**Phase3:** The third focus concerns the more general development of clinical decision support systems (CDSSs) that, based on the collected patient data from physical exams, laboratory results etc., can provide early warnings to the healthcare professionals. These systems range from being purely rule-based to using techniques such as artificial neural networks, support vector machines and decision trees to create models suitable for sepsis diagnosis

**Phase4:** A web-based support system for sepsis treatment is presented in, where the healthcare practitioners are guided through the sepsis diagnosis process with the help of rule-based and data-driven logic algorithms, providing patient adapted treatment instructions

**2.4 Features of the Project**

The project is analysis-based project hence the features of this project are as:

1. Analysis on real time data set (Original dataset)

2. Creating an Expert System and making decision

3. Website and mobile-app interface for the features

**2.5 Methodology Used**

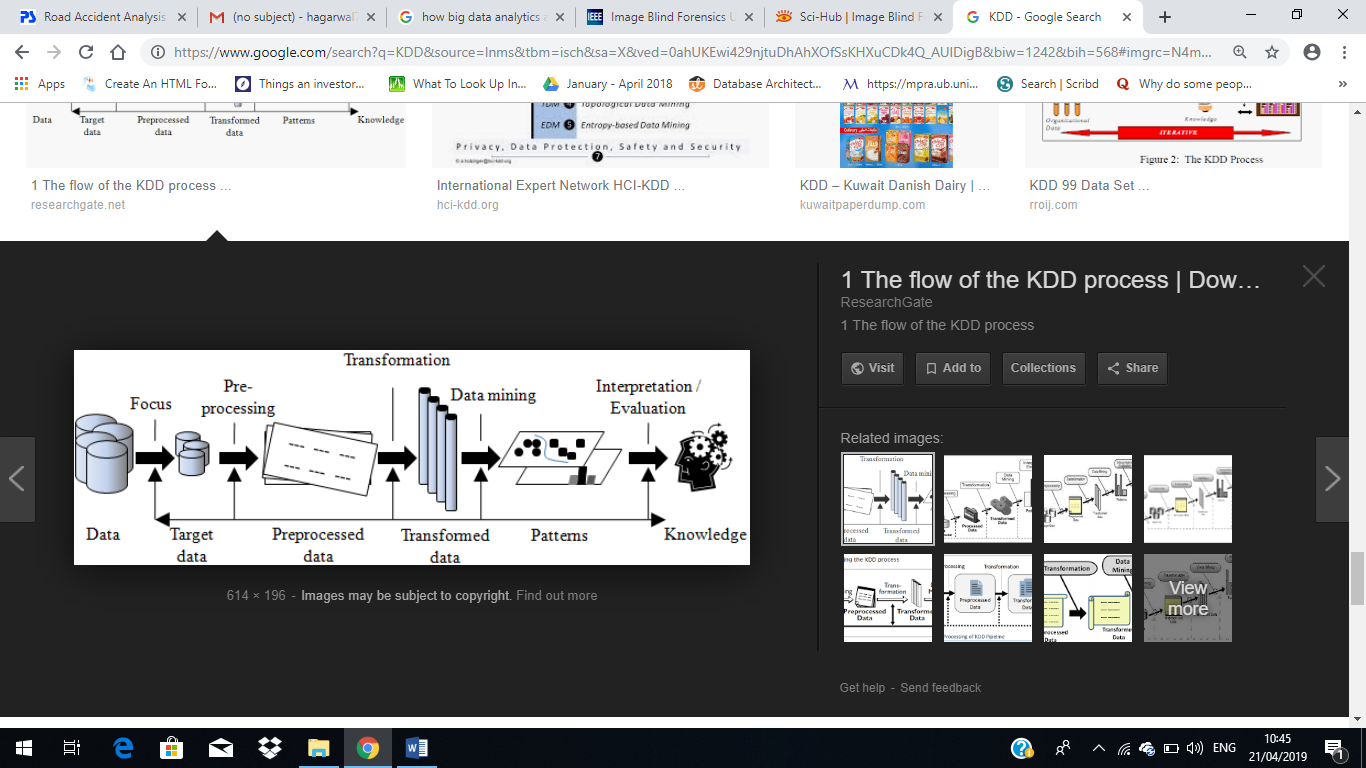


Fig.2.5.1: KDD process

For implementing the given problem definition, the classical KDD (Knowledge Discovery in Database) process which refers to non- trivial extraction of implicit, previously unknown and potentially useful information from the data stored in databases. This process is going to be used which involves the following:

1. Data collection: it’s the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Data for this project would be collected from various databases

1. BIDMC
2. UCSF
3. PubMed
4. Embase
5. Scopus
6. Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.
7. Cleaning in case of Missing values.
8. Cleaning noisy data, where noise is a random or variance error.
9. Cleaning with Data discrepancy detection and Data transformation tools.
10. Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).
11. Data integration using Data Migration tools.
12. Data integration using Data Synchronization tools.
13. Data integration using ETL(Extract-Load-Transformation) process.
14. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

Through our rigorous literature survey, we found heartbeat, Respiratory rate, Temperature, Systolic blood pressure, Diastolic blood pressure, arterial blood pressure, pulse pressure, age, white blood cell count are the important features.

1. Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two-step process:

* 1. Data Mapping: Assigning elements from source base to destination to capture transformations.
  2. Code generation: Creation of the actual transformation program

1. Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   1. Transforms task relevant data into patterns.
   2. Decides purpose of model using classification or characterization.
2. Pattern Evaluation: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
3. Find interestingness score of each pattern.
4. Uses summarization and Visualization to make data understandable by user.
5. Knowledge Representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
6. Generate reports.
7. Generate tables.
8. Generate discriminant rules, classification rules, characterization rules, etc.

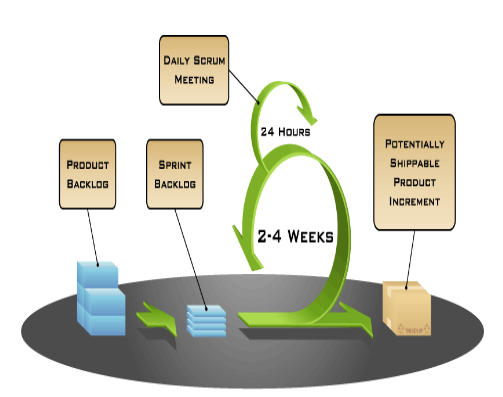


Fig.2.5.2: SCRUM

Scrum is an innovative approach to getting work done in efficient way. It is iterative & incremental agile software development method. These iterations are time boxed with various iterations & each iteration is called Sprint. The Sprint is basically 2-4 week long & each sprint requires sprint planning estimation. According to latest surveys Scrum is the most popular agile project management methodology in software development.

Scrum is ideally used where highly emergent or rapidly changing requirements. Scrum is basically worked on a self-organizing, cross-functional team. In the overall scrum team there is no team leader who assign the task to team rather whole scrum members work as a team & they decides the task on which they will work on. Also the problem will be resolve by team.

**Each Agile Development Scrum team having three core scrum roles:** Product Owner, Scrum Master & The Team.

**1) Product Owner:** The Product Owner is the person who represents the stakeholders and is the voice of the customer. Product owner writes the User Stories, ordered priorities and add in the Product Backlog. It is recommended that Agile Scrum Master should not mix with Product Owner.

**2) Scrum Master:** The Scrum-Master is a facilitator, team leader who ensures that the team adheres to its chosen process and removes blocking issues to deliver the sprint deliverable/goal. Scrum Master is not a team leader but act as a shield for the team from external interference’s & also removes barriers.

**3) The Team:** The scrum development team is generally size of 5-9 peoples with self-organizing and cross-functional skills who do actual work like Analysis, Design, Development, Testing, Documentation etc.

**CHAPTER 3**

**ANALYSIS & PLANNING**

**3.1 Introduction**

Getting a clear idea of the project title and doing research on it we will get our definition and after that then we will first create the Literature Survey of the project and do the whole documentation. After analysis, we will first study about it and do some research on it for our better understanding of the project and also get a rough picture about what would be our problem definition for the particular project.

**3.2 Feasibility Study**

**1. Technical Feasibility** - This assessment focuses on the technical resources available to the organization. It helps organizations determine whether the technical resources meet capacity and whether the technical team is capable of converting the ideas into working systems. It also involves the evaluation of the hardware, software, and other technical requirements of the proposed system.

**2. Economic Feasibility** - This assessment typically involves a cost/ benefits analysis of the project, helping organizations determine the viability, cost, and benefits associated with a project before financial resources are allocated. It also serves as an independent project assessment and enhances project credibility— helping decision makers determine the positive economic benefits to the organization that the proposed project will provide.

**3. Legal Feasibility** - This assessment investigates whether any aspect of the proposed project conflicts with legal requirements like zoning laws, data protection acts, or social media laws. Let’s say an organization wants to construct a new office building in a specific location. A feasibility study might rev

eal the organization’s ideal location isn’t zoned for that type of business. That organization has just saved considerable time and effort by learning that their project was not feasible right from the beginning.

**4. Operational Feasibility** - This assessment involves undertaking a study to analyze and determine whether—and how well—the organization’s needs can be met by completing the project. Operational feasibility studies also analyze how a project plan satisfies the requirements identified in the requirements analysis phase of system development.

**5. Scheduling Feasibility** - This assessment is the most important for project success; after all, a project will fail if not completed on time. In scheduling feasibility, an organization estimates how much time the project will take to complete.

**3.3 Project Planning**

**Phases of the Project:S**

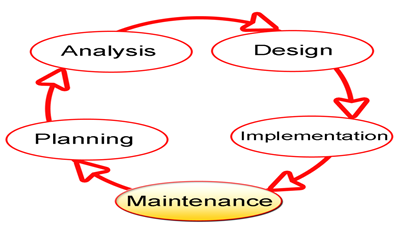


Fig.3.3.1: Phases of project development

1. **System Planning**

The planning phase is the fundamental process of understanding why an information system should be built and determining how the project team will go about building it. It has two steps:

* During project initiation, the system's business value to the organization is identified how will it lower costs or increase revenues? Most ideas for new systems come from outside the IS area (from the marketing department, accounting department, etc.) in the form of a system request. A system request presents a brief summary of a business need, and it explains how a system that supports the need will create business value.
* Once the project is approved, it enters project management. During project management, the project manager creates a work plan, staffs the project, and puts techniques in place to help the project team control and direct the project through the entire SDLC.

1. **System Analysis**

The analysis phase answers the questions of who will use the system, what the system will do, and where and when it will be used. During this phase, the project team investigates any current system(s), identifies improvement. This phase has three steps:

* 1. An analysis strategy is developed to guide the project team's efforts. Such a strategy usually includes a study of the current system (called the as-is system) and its problems, and envisioning ways to design a new system (called the to-be system).
  2. The next step is requirements gathering (e.g., through interviews, group work-shops, or questionnaires). The analysis of this information—in conjunction with input from the project sponsor and many other people—leads to the development of a concept for a new system.
  3. The analyses, system concept, and models are combined into a document called the system proposal, which is presented to the project sponsor and other key decision makers (e.g., members of the approval committee) who will decide whether the project should continue to move forward.

1. **System Design**

The design phase comes after a good understanding of customer’s requirements, this phase defines the elements of a system, the components, the security level, modules, architecture and the different interfaces and type of data that goes through the system. The design phase decides how the system will operate in terms of the hardware, software, and network infrastructure that will be in place; the user interface, forms, and reports that will be used; and the specific programs, databases, and files that will be needed. The design phase has four steps:

1. The design strategy must be determined. This clarifies whether the system will be developed by the company's own programmers, whether its development will be outsourced to another firm (usually a consulting firm), or whether the company will buy an existing software package.
2. This leads to the development of the basic architecture design for the system that describes the hardware, software, and network infrastructure that will be used.
3. The database and file specifications are developed. These define exactly what data will be stored and where they will be stored.
4. The analyst team develops the program design, which defines the programs that need to be written and exactly what each program will do.

**4. Implementation and Deployment**

This phase comes after a complete understanding of system requirements and specifications, it’s the actual construction process after having a complete and illustrated design for the requested system.

In the Software Development Life Cycle, the actual code is written here, and if the system contains hardware, then the implementation phase will contain configuration and fine-tuning for the hardware to meet certain requirements and functions.

In this phase, the system is ready to be deployed and installed in customer’s premises, ready to become running, live and productive, training may be required for end users to make sure they know how to use the system and to get familiar with it, the implementation phase may take a long time and that depends on the complexity of the system and the solution it presents.

**5. System Maintenance**

In this phase, periodic maintenance for the system will be carried out to make sure that the system won’t become obsolete, this will include replacing the old hardware and continuously evaluating system’s performance, it also includes providing latest updates for certain components to make sure it meets the right standards and the latest technologies to face current security threats.

These are the main six phases of the System Development Life Cycle, and it’s an iterative process for each project.

**3.4 Scheduling**

**Gantt chart:**

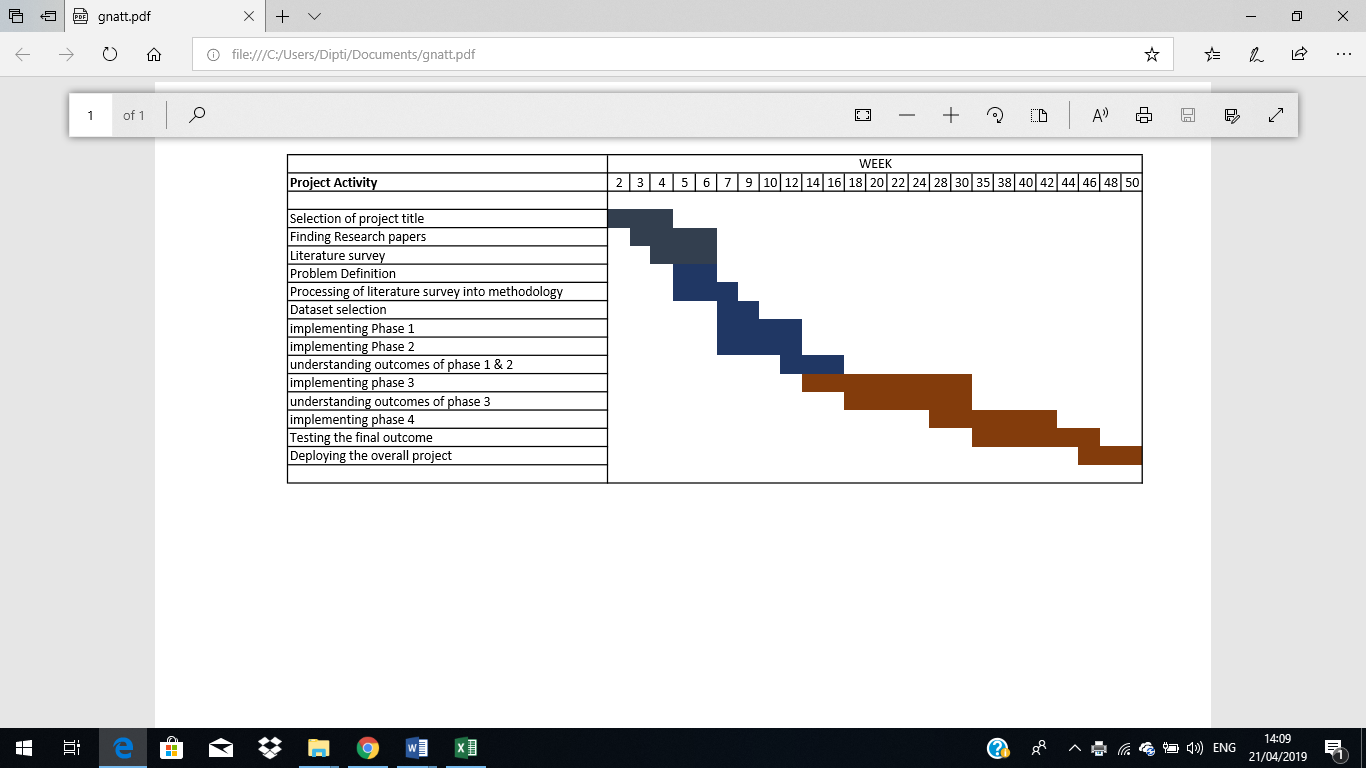


Fig.3.4.1: Gnatt Chart

It shows the representation of each day of the work from the Selection of Project Title to its Testing.

In this after understanding the topic the project feasibility was analyzed by performing different types of feasibility studies and by also planning the project tools, their project schedule, timeline charts, etc. Feasibility study will help in better understanding the various feasibilities associated with the project and helping to make the correct decisions and completing the project within the schedule, budget, etc.

The tools were specifically identified in this chapter stating which technology can be feasible and how conveniently the project can be completed. This helps to understand the technology and tools that can be used for the project. The Gantt chart helps us to track the project and see the schedule of the project and to see if the project is on the right track and on schedule and not behind the deadline

**CHAPTER 4**

**DESIGN & IMPLEMENTATION**

**4.1 Model Requirements**

The Model requirements for our project are as follows:

1. **Dataset:** Large data sets are first sorted, then patterns are identified and relationships are established to perform data analysis and solve problems.
2. **Anaconda:** Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. Package versions are managed by the package management system conda.
3. **Algorithms:** In today’s world of “big data”, a large database is becoming a norm. Just imagine a database with many terabytes. As facebook alone crunches 600 terabytes of new data every single day. Also, the primary challenge of big data is how to make sense of it. Moreover, the sheer volume is not the only problem. Also, big data need to diverse, unstructured and fast changing. Consider audio and video data, social media posts, 3D data or geospatial data. This kind of data is not easily categorized or organized. Further, to meet this challenge, a range of automatic methods for extracting information.

**4.2 Block Diagram**

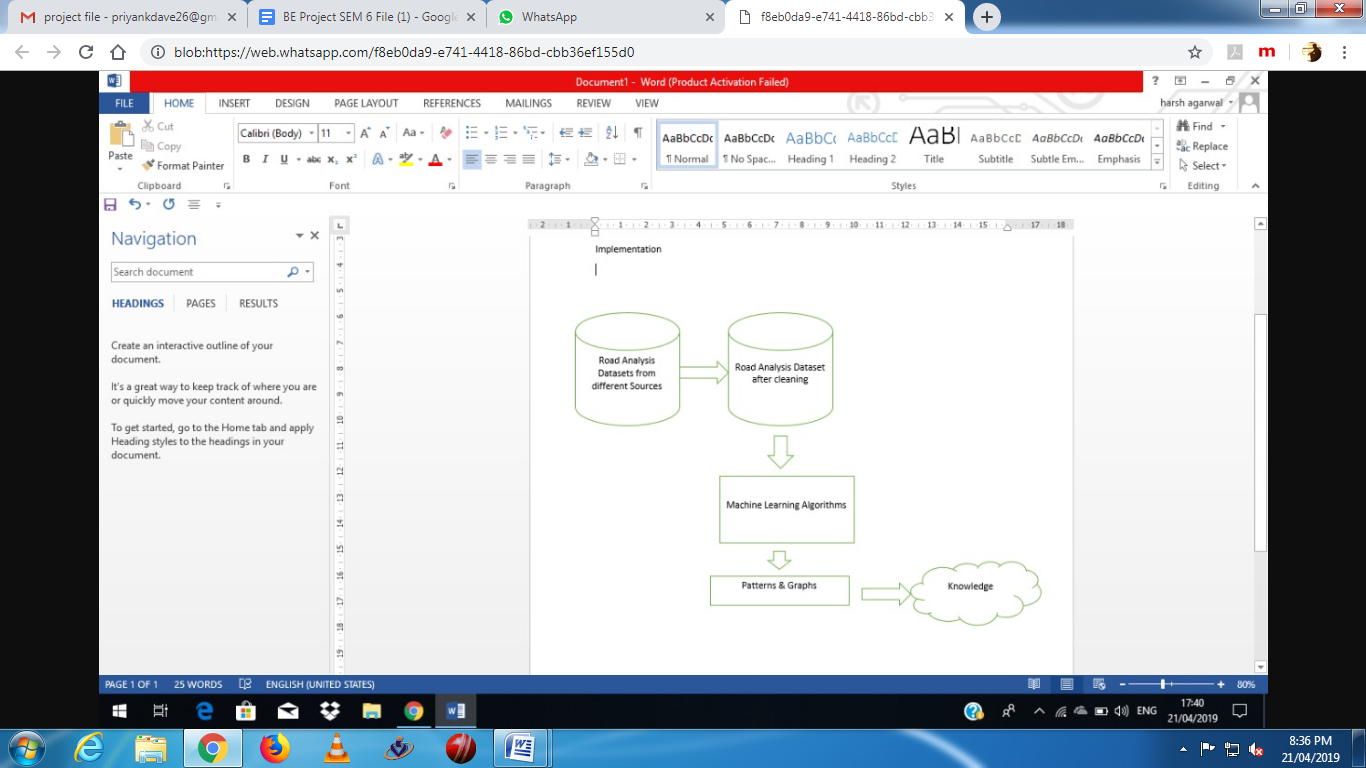
****

Fig.: Block diagram

**4.3 Flow Chart**

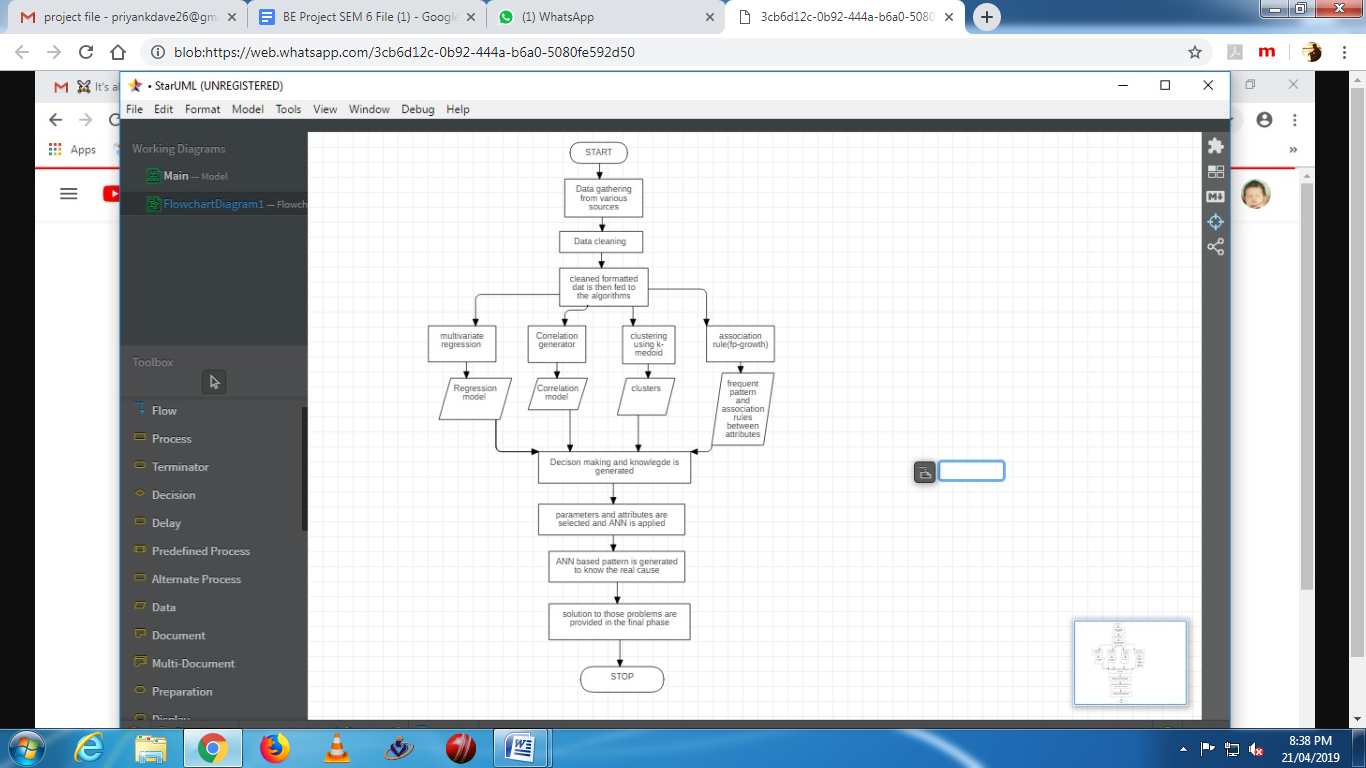
****

Fig.4.3.1: flowchart

**4.4 Database Screenshot**

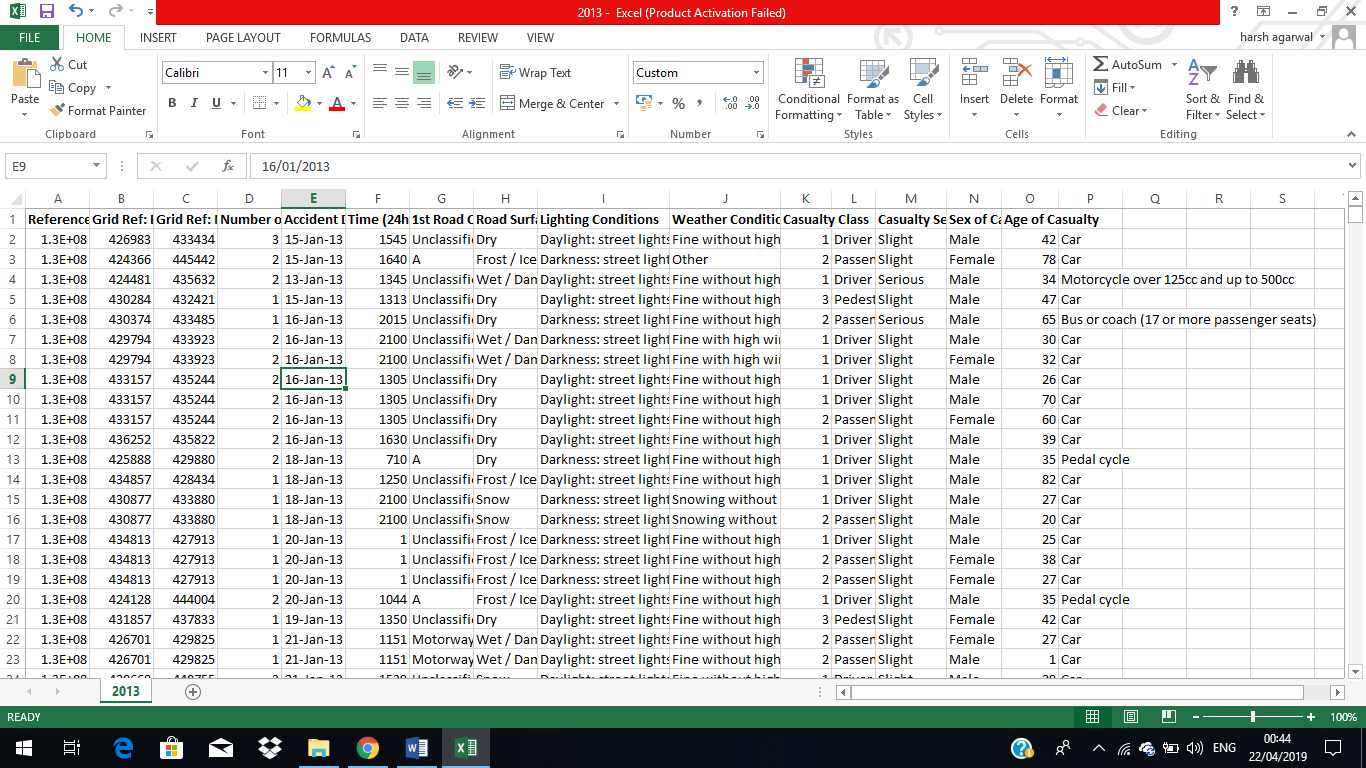


Fig.: Database Screenshot

Dataset size: 2420\*14

Attributes: 14

**CHAPTER 5**

**RESULTS & DISCUSSION**

**5.1 Results**

**5.2 Future Scope**

From the above comparative analysis of the various credit card fraud detection techniques it is clear that Artificial Neural Networks performs best in this scenario. But the drawbacks of Artificial Neural Networks is that they are very expensive to train and can be easily over trained. In order to minimize their expense we need to create a hybrid of neural network with some optimisation technique. Optimisation techniques that could be successfully paired with Neural Network are Genetic Algorithm, Artificial Immune System, Case Based Reasoning and any other similar optimisation technique. Genetic Algorithm helps by selecting the optimised weight of the edges in neural network. Artificial Immune System reduces the cost by eliminating the weights that cause the maximum error and Case Based Reasoning first tries to predict the outcome on the basis of a direct match with the user’s profile.

**CHAPTER 6**

**CONCLUSION**

**6. Conclusion**

Although there are several fraud detection techniques available today but none is able to detect all frauds completely when they are actually happening, they usually detect it after the fraud has been committed. This happens because a very minuscule number of transactions from the total transactions are actually fraudulent in nature. So we need a technology that can detect the fraudulent transaction when it is taking place so that it can be stopped then and there and that too in a minimum cost. So the major task of today is to build an accurate, precise and fast detecting fraud detection system for credit card frauds that can detect not only frauds happening over the internet like phishing and site cloning but also tampering with the credit card itself i.e. it signals an alarm when the tampered credit card is being used.

**References**

|  |
| --- |
| [1] KhyatiChaudhary, JyotiYadav, BhawnaMallick, “ A review of Fraud Detection Techniques: Credit Card”, International Journal of Computer Applications Volume 45– No.1 2012.  [2] Michael Edward Edge, Pedro R, Falcone Sampaio, “A survey of signature based methods for financial fraud detection”, journal of computers and security, Vol. 28, pp 3 8 1 – 3 9 4, 2009.  [3] Linda Delamaire, Hussein Abdou, John Pointon, “Credit card fraud and detection techniques: a review”, Banks and Bank Systems, Volume 4, Issue 2, 2009.  [4] Salvatore J. Stolfo, David W. Fan, Wenke Lee and Andreas L. Prodromidis; "Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results"; Department of Computer ScienceColumbia University; 1997.  [5] Maes S. Tuyls K. Vanschoenwinkel B. and Manderick B.; "Credit Card Fraud Detection Using Bayesian and Neural Networks"; Vrije University Brussel – Belgium; 2002.  [6] Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2000.  [7] Salvatore J. Stolfo, Wei Fan, Wenke Lee and Andreas L. Prodromidis; "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project"; 0-7695-0490-6/99, 1999 IEEE.  [8] Soltani, N., Akbari, M.K., SargolzaeiJavan, M., “A new user-based model for credit card fraud detection based on artificial immune system,” Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on., IEEE, pp. 029-033, 2012.  [9] S. Ghosh and D. L. Reilly, “Credit card fraud detection with a neural-network”, Proceedings of the 27th Annual Conference on System Science, Volume 3: Information Systems: DSS/ KnowledgeBased Systems, pages 621-630, 1994. IEEE Computer Society Press.  [10] MasoumehZareapoor, Seeja.K.R, M.Afshar.Alam, ”Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria”, International Journal of Computer Applications (0975 – 8887) Volume 52– No.3, 2012.  [11] Fraud Brief – AVS and CVM, Clear Commerce Corporation, 2003, http://www.clearcommerce.com. [12]All points protection: One sure strategy to control fraud, Fair Isaac, http://www.fairisaac.com, 2007. |
|  |