International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 29 Issue 4 – MAY 2022.

SOFTWARE BASED PREDICTION OF LIVER DISEASE USING MACHINE LEARNING TECHNIQUES

A.J.R. Bewin[1], N. Sathishkumar[1], B. Sureshkannan[1], A.Velmurugan [1], Mrs.S.T.Shenbagavalli[2]

Students-Department of CSE-PSNA,Dindigul[1]

Assistant Professeror-Department of CSE-PSNA, Dindigul[2]

***Abstract*—Today's Health care is a very important aspect for every human being, hence there is a need to provide medical services that are easily available to everyone. The liver is the largest organ of the body and it is essential for digesting food and releasing the toxic substance of the body. Liver disorders have increased r and it is considered to be a very fatal disease in many countries. Since the lifestyle of the human changes, their cuisines and eating habit also changes. But this change is not easily accepted by the internal organs. Especially digestion is important process which gets affected. All the blood leaving into stomach and intestines passes through the liver and its helps in excretion of cholesterol, hormone, bilirubin and drugs. They also do metabolism activity which involves protein, carbohydrate and fats. Yearly 2 million people dies due to the Liver disease according to National Library of Medicine, USA. So, it is necessary to predict the possibility of the disease before it gets vigorous. The main focus of the proposed project is to predict the liver disease based on Machine Learning approach using classification algorithms. Machine learning has made a great impact on the biomedical field for liver disease prediction and diagnosis. Machine learning offers a guarantee for improving the detection and prediction of disease that has been made an interest in the biomedical sector and they also increase the objectivity of the decision-making process. The main aspect is to predict the results more efficiently and reduce the cost of diagnosis in the medical field. Therefore, we used different classification techniques for the classification of patients have the liver disease.**

Keywords: Machine learning, Hypertext Mark-up Language, Cascade Style Sheet, Support

I. INTRODUCTION

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.[1]

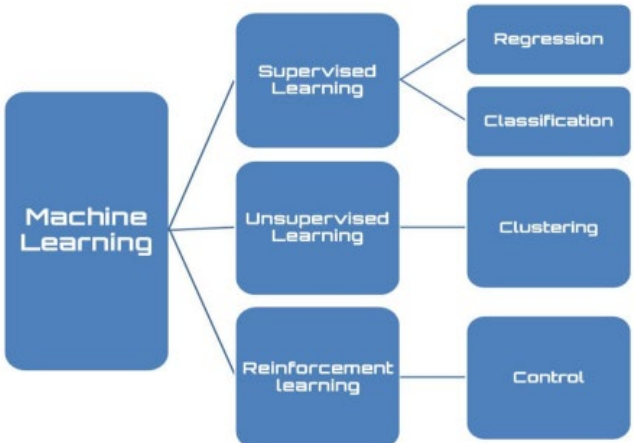


Figure 1.Machine Learning.

Liver disease (also called hepatic disease) is a type of damage to or disease of the liver. The liver is the largest organ in the body. Amongst its 500+ roles, the liver is responsible for food processing, energy storage, blood filtration, and immune response. Specifically, the liver contributes by secreting bile for lipid breakdown, storing excess glucose as glycogen, and removing bacteria and toxins from the blood. Whenever the course of the problem lasts long, chronic liver disease ensues. Therefore, this makes the way to develop a system that helps in prediction of liver disease.

Proposed model should able to get the data from user

and compare that with existing model and should show the risk of affecting disease should be shown.With help of the prediction, the patient should be able to advance their medication in order to avoid getting disease at severe level.

II. Related works

Rashid Naseem et al. proposed a Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome. This study compares ten classifiers including A1DE, NB, MLP, SVM, KNN, CHIRP, CDT, Forest-PA, J48, and RF to find the optimal solution for early and accurate prediction of liver disease. The datasets utilized in this study are taken from the UCI ML repository and the GitHub repository. The outcomes are assessed via RMSE, RRSE, recall, specificity,precision, G-measure, F-measure, MCC, and accuracy. The exploratory outcomes show a better consequence of RF utilizing the UCI dataset. Assessing RF using RMSE and RRSE, the outcomes are 0.4328 and 87.6766, while the accuracy of RF is 72.1739% that is also better than other employed classifiers. However, utilizing the GitHub dataset, SVM beats other employed techniques in terms of increasing accuracy up to 71.3551%.

Baddigam Jaya Krishna Reddy et al. proposed a Diagnosis of Liver Disease using Machine Learning Model. The methods of Support Vector Machines (SVM), Decision Tree (DT) and Random Forest (RF) is proposed to predict liver disease with better precision, accuracy and reliability. Hence, the aim of this project is to investigate the data mining algorithm to predict liver disease on imbalanced data through random sampling. Results are compared and analysed based on accuracy and ROC index. K-Nearest Neighbour (k-NN) outperforms the other algorithms such as Logistic Regression, AutoNeural and Random Forest with the accuracy of 99.794%.

Jagdeep Singha et al. proposed a Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. The main focus is to predict the liver disease based on a software engineering approach using classification and feature selection technique. The implementation of proposed work is done on Indian Liver Patient Dataset (ILPD) from the University of California, Irvine database. Proposed work focuses on the development of the software that will help in the prediction of the level diseases based upon the various symptoms.

Vijay Panwar, et al. proposed a Review Of Liver Disease Prediction Using Machine Learning Algorithm. The performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. In this work, they used five algorithms Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, and specificity. We found the accuracy 74%, 72%, 72%, 71%, and 57% for SVM,DT,RF,LR and NB. The analysis result shown the SVM achieved the highest accuracy. Moreover, our present study mainly focused on the use of clinical data for liver disease prediction and explores different ways of representing such data through our analysis.

Muktevi Srivenkatesh et al. Proposed a Performance Evolution of Different Machine Learning Algorithms for Prediction of Liver Disease. Liver sickness might be distinguished with incalculable order systems, and these have been classified the utilization forecast of a number highlights and classifier blends. In this investigation, they applied five sort of classifiers that is Naïve Bayes, logistic regression, support vector machines, Random Forest, K Nearest Neighbour for the examination of liver malady. The classification exhibitions are assessed with 5 distinctive by and large execution measurements, i.e., precision, kappa, Mean absolute error (MAE), Root mean square error (RMSE), and F measures. The objective of this query work is to foresee liver infection with different machine learning and pick most efficient algorithm.

Fahad Mostafa et al. The purpose of this study was to extract significant predictors for liver disease from the medical analysis of 615 humans using ML algorithms. Data visualizations were implemented to reveal significant findings such as missing values. Multiple imputations by chained equations (MICEs) were applied to generate missing data points, and principal component analysis (PCA) was used to reduce the dimensionality. Variable importance ranking using the Gini index was implemented to verify significant predictors obtained from the PCA. Training data (ntrain=399) for learning and testing data (ntest=216) in the ML methods were used for predicting classifications. The study compared binary classifier machine learning algorithms (i.e., artificial neural network, random forest (RF), and support vector machine), which were utilized on a published liver disease data set to classify individuals with liver diseases, which will allow health professionals to make a better diagnosis. The synthetic minority oversampling technique was applied to oversample the minority class to regulate overfitting problems. The RF significantly contributed ($p<0.001$) to a higher accuracy score of 98.14% compared to the other methods. Thus, this suggests that ML methods predict liver disease by incorporating the risk factors, which may improve the inference-based diagnosis of patients.

III. Proposed System

The proposed system will predict the risk factor of the HCC affected patients using Machine Learning. The first step is the extraction of data. Data set from Kaggle where biological data of 165 people are collected. Since the data is collected on real time, data preprocessing is needed. In preprocessing we are eliminating null values and inconsistent data and we will get dataset which will used be for processing. Using DBN algorithm, we are getting best features for processing. By using this feature we can able to

build best classifier model. After getting dataset with best feature, we want the build suitableclassifier model which will give best accuracy. For our use case, we are using SVM algorithm.

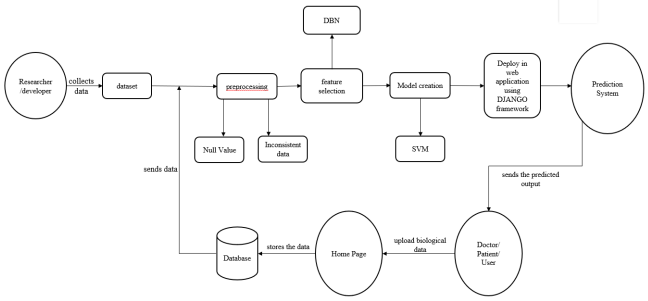
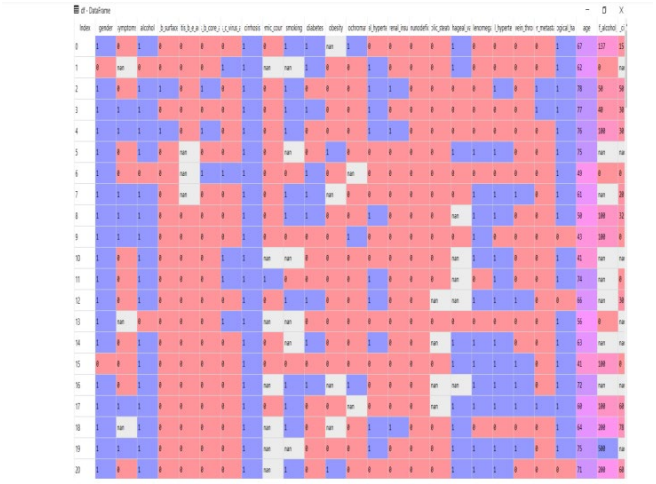


Figure 2.Architectural Diagram.

IV. WORKING AND IMPLEMENTATION

A. *Data Collection*

To achieve the goal, Data Engineering is the first step. Data Engineering consists of two processes, they are Data Collection and Data Pre-processing. Data Collection will be collected with meaningful parameters like age, blood test and so on.[3]

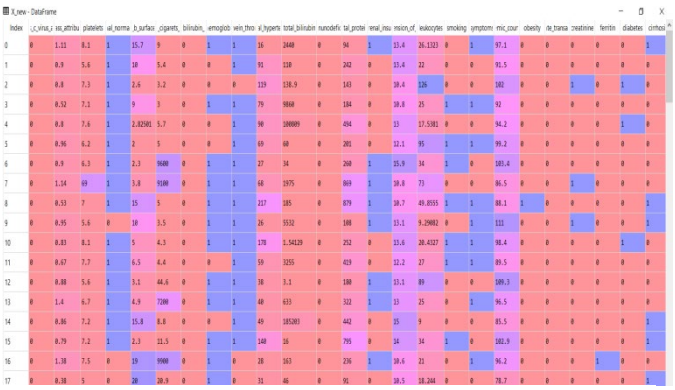


Data Pre-Processing: Null value count in each feature input

```
In [3]: print("Null values colunt for each column")
   ...:
   ...: data.isnull().sum(axis=0)
Null values colunt for each column
Out[3]:
gender                          0
symptoms                       18
alcohol                         0
hepatitis_b_surface_antigen    17
hepatitis_b_e_antigen          39
hepatitis_b_core_antibody      24
hepatitis_c_virus_antibody      9
cirrhosis                       0
endemic_countries              39
smoking                        41
diabetes                        3
obesity                        10
hemochromatosis                23
arterial_hypertension           3
chronic_renal_insufficiency     2
human_immunodeficiency_virus   14
nonalcoholic_steatohepatitis   22
esophageal_varices             52
splenomegaly                   15
portal_hypertension            11
portal_vein_thrombosis          3
liver_metastasis                4
radiological_hallmark           2
age                             0
grams_of_alcohol_per_day       48
packs_of_cigarets_per_year     53
performance_status              0
encephalopathy_degree           1
ascites_degree                  2
international_normalised_ratio  4
```

B. *Data Pre-processing*

Collected data will be pre-processed which means encoding the categorical information in the data. Dropping unwanted parameters, scaling the parameter values to achieve normal distribution (Zero mean and Standard Deviation as one), handling missing values and so on. Here Data set from Kaggle where biological data of 165 people are collected[4]



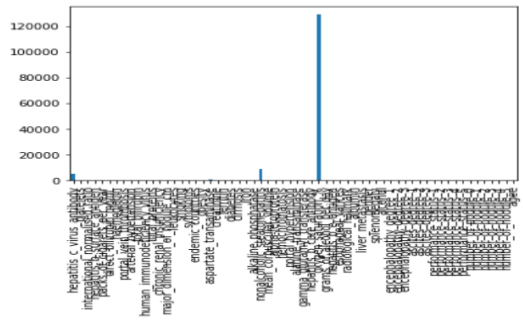
```
In [5]: print("After preprocessing : filled null values")
   ...: df2.isnull().sum()
After preprocessing : filled null values
Out[5]:
hepatitis_c_virus_antibody        0
class_attribute                   0
platelets                         0
international_normalised_ratio    0
hepatitis_b_surface_antigen       0
                                 ..
number_of_nodule_2                0
number_of_nodule_3                0
number_of_nodule_4                0
number_of_nodule_5                0
agee                              0
Length: 63, dtype: int64

In [6]:
```

C. *Feature Engineering*

After the Data Engineering process, Feature Engineering will be done. Feature Engineering is an important step to predict our output. The advantage of Feature Engineering is minimizing the parameter. For example, if our whole dataset contains 10 parameters, after feature engineering only three-parameter enough to predict the output with high efficiency. Feature Engineering based on correlation, co-variance, co-linearity and etc. Feature Engineering has

many algorithms to predict correct correlated parameters.[5-6]

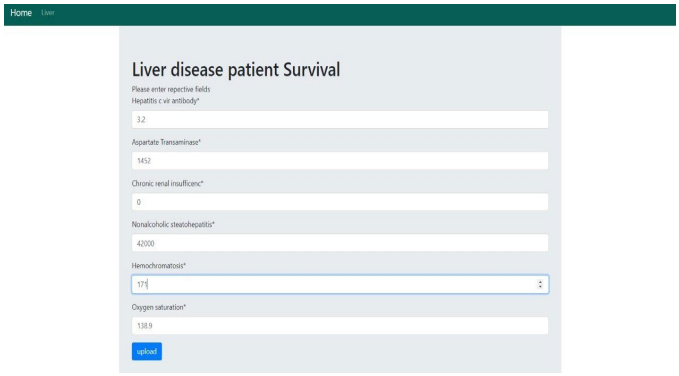


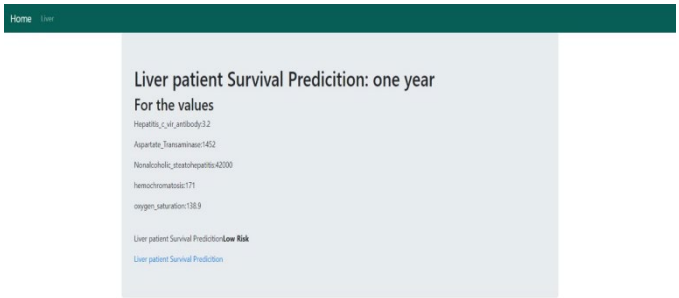
D. Building Model

After getting dataset with best feature, we want the build suitable model which will give better accuracy. For our use case, we are using SVM algorithm. Among three algorithms, we are selecting one algorithm for our model which gives high accuracy.[6-8]

E. Web Deployment

The trained model will be saved and loaded for web development. With the help of a built model and with a selected feature we can able to predict the employee resignation. Web development will have an input variable of selected features, by submitting the answer of the selected feature, the prediction will be done.





V. TESTING

Since the error in the software can be injured at any stage. So, we have carry out the testing process at different levels during the development. The basic levels of testing are,

- Unit Testing
- Integration Testing
- Validation Testing
- Functional Testing
- Structural Testing

A. Unit Testing

Unit testing was used to test individual units in the system and ensure that they operate correctly. Alternate logic analysis and screen validations were tested in this to ensure optimum efficiency in the system. The procedures and functions used and their association with data were tested.

B. Integration Testing

This testing process focuses on identifying the interfaces between components and their functionality. The bottom up approach was adopted during this testing. Low-level modules are integrated and combined as a cluster before testing. This allowed identifying any wrong linkages or parameters passing early in the development process as it just can be passed in the set of data and checked if the result returned is an accepted one.

C. Validation Testing

Software testing and validation is achieved through a series of block box tests that demonstrate conformity with requirements. A test procedure defines specific test cases that will be used to demonstrate conformity with requirements. Both, the plan and the procedure are designed to ensure that all functional requirements are achieved, documentation is correct and other requirements are met. After each validation test case has been conducted, one of the two possible conditions exists.

D. Functional Testing

Functional testing, also known as block box or closed box testing, is normally applied to HDL (High-Level Data Link) code that operates concurrently and concentrates on checking the interaction between modules, blocks or functional boundaries. The objective here is to ensure that `correct results" are obtained when `good inputs" are applied operates in a predictable manner. Functional testing can therefore be considered as concentrating on checking that the data paths operate correctly. The coverage measurements that fall into this category are toggle, triggering, and signal trace coverage.

E. Structural Testing

Structural testing, are known as white box or open box testing, is normally applied to sequential HDL (High-Level Data Link) code and concentrates on checking that all executable statements within each module have been exercised and the corresponding branches and paths through that module have been covered. If there is a section of HDL code that has never been exercised then there is a high possibility that it could contain an error that will remain undetected.

VI. DRIVING TEST CASES

A test case is a set of conditions or variables under which a tester will determine if a requirement upon an application is partially or fully satisfied. The types of testing that are to be carried out on the system is as follows.

Test Case no	Description	Pre condition	Pass/Fail	Expected Results
1	Check the homepage is available at http://127.0.0.1:8000/	Page should be available	Pass	Home page should appear
2	Check feature selected attribute is selected	Attributes should be present	Pass	All the attributes are present
3	Valid data should be entered	No data other than number should be entered	Pass	Alert should be displayed when wrong datatype is entered
4	Once upload button is clicked, data must to be send to prediction system	Prediction display must to be appear	Pass	Prediction display page should appear
5	Check whether correct output is shown	Output should be generated correctly	Pass	Output should be generated

VII. CONCLUSION

The HCC affected person's risk factor was classified with Support Vector Machine. This was achieved with feature selection method select-K parameter with chi square. The effective five features were selected from 50 features using feature selection method. The result achieved was 95% accuracy. The trained model with SVM for 5 feature input are able to predict the low risk or high risk. Advantage of using feature selection has eliminated the unwanted feature which may increase the blood test cost of the person.

In the proposed work, different classifiers were implemented on liver patient diseases dataset to predict liver diseases based on developed software. Dataset was processed and implemented using feature selection techniques. The results of the proposed work were compared using feature selection and without using feature selection techniques after the implementation of different classifiers in terms of execution time and accuracy.

The best result was achieved using Logistic Regression classifier with feature selection techniques and execution time of different classifiers was decreased after the implementation of feature selection technique. Finally, liver disease prediction Software (LDPS) is developed using concept of software engineering life cycle

VIII. FUTURE WORK

The application developed is simple prototype to explain the basic functionalities of the upcoming application. In the upcoming release following features will be added

- Prediction of liver disease can be added to hospital management system.
- Image processing system can be added.
- Prescription can be suggested based on the risk.

IX. REFERENCES

[1] Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome Rashid Naseem, Bilal Khan, M. A. Shah, Karzan Wakil, Atif Khan, Wael Alosaimi, M. I. Uddin, Badar Alouffi Published on 2020 in "National Library of medicine, USA".

[2] Diagnosis of Liver Disease using Machine Learning Models A. Sivasangari, Baddigam Jaya Krishna Reddy, Annamareddy Kiran, P. Ajitha Published 2020 on "IEEE xplore".

[3] A.N.Arbainand, B.Y.P.Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data, published in " International Journal of Data Science and Analytics, vol. 1, on 2019.

[4] Jagdeep Singha, Sachin Baggab, Ranjodh Kaur Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques International Conference on Computational Intelligence and Data Science (ICCIDS 2019).

[5] Binish Khan Piyush Kumar Shukla Manish Kumar Ahirwar Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING · July 2019.

[6] A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data "Ain Najwa Arba, B. P. Balakrishnan Published 2019 published in International Journal of Data Science and Advanced Analytics (ISSN 2563-4429).

[7] T. Choudhury, and A. Thakral. (2019), "Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques" published in IEEE International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 163-168.

[8] I. Arshad, C. Dutta, T. Choudhury, and A. Thakral. (2019), "Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques." Published in IEEE xplore.