JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANALYSIS AND PREDICTION OF LIVER DISEASE USING CNN AND KNN

[1] Gudiwaka Vijayalakshmi ,[2]Rongala Rajesh
[1]Assistant Professor,[2]MCA 2nd year,
[1]Computer Science and Engineering, [2]Master of Computer Applications,
Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

ABSTRACT
The rapid rise in the prevalence of obesity and an unhealthy lifestyle eventually reflects the incidence and frequency of liver-related disorders in the general population In this effort, patient data sets are analysed for their predictability of having a liver disease based solely on a widely studied classification model Because pre-existing processes for analysing patient and classifier data exist, the more significant aspect here is to forecast the same decisive outcome with a better rate of accuracy This procedure is divided into five main stages To begin, the min-max algorithm is applied to the original liver patient data set obtained from the UCI repository Using PSO feature selection, key properties are delineated in the second step In the third phase, classification algorithms are used for comparative analysis and categorization The fourth phase is the accuracy calculation It makes use of the Root Mean Square Value and the Root Error Value The fifth and final phase is the evaluation phase With an accuracy rate of 95.04%, the CNN algorithm is considered to be a superior performing algorithm when it comes to feature selection.

***Keywords*:**Obesity,Liverdisease,Patient datasets,classification model,PSO featureselection, classification algorithms, accuracy calculation, Root Mean Square, Root Error Value, CNN algorithm, feature selection, accuracy rate, predictive models, systematic approach**.**

I.INTRODUCTION

Machine learning is applied in a variety of industries. The healthcare industry is no different. Machine learning can be very beneficial in predicting the existence of diseases such as locomotor disorders, cardiovascular diseases, liver diseases, and others. Such evidence, if expected, can give clinicians with valuable knowledge, allowing them to tailor their treatment plans and diagnoses[1].The digital technological revolution is demonstrating its potential for disruptive innovation[2].With the advancement of medical technologies such as nanotechnology and genetics, the sky appears to be the limit when it comes to imagining the various ways to utilise the enormous potential of the digital marketing era, ineffective prognosis, diagnosis, treatment, and healthcare monitoring.A significant amount of data is dealt with on a regular basis, in relation to the passage of each instance of a medical procedure working[3]. These data sets could be inferential, referential, or raw enough to be conclusive of further relevant sets of useful medical information[4]. This information comes from a variety of sources and is used in a variety of ways. They can be used to predict, diagnose, and treat diseases. The investigation of the same could hasten the pace of research initiatives in the same environment[5]. It may aid in statistical conclusions when it comes to projecting various patterns that may aid in the overall process[6]. Classification algorithms are widely used in data mining for medical diagnosis and disease prediction[7].The liver is the second biggest internal organ in the human body, playing an important role in metabolism and performing various crucial activities such as red blood cell decomposition, among others. It weighs approximately three pounds. The liver is responsible for several vital tasks such as digestion, metabolism, immunity, and nutrient storage inside the body[8]. These functions distinguish the liver as a vital organ; without it, body tissues would perish due to a lack of energy and nutrition[9]. Traditionally, liver disease has been clinically diagnosed by analysing enzyme levels in the blood[10]. In this study, a mix of Naive Bayes and Support Vector Machine (SVM) classifier methods is utilised to predict liver illness[11].

II.EXISTING SYSTEM

Prediction systems for liver illness have played an essential role in people's lives, and many scholars believe it is an important topic. Although the forecast findings are encouraging, these old methods are still far from being highly exact and efficient. Existing methods are simple and functional, but they are extremely vulnerable to interruption[12].Furthermore, cutting-edge approaches only use one algorithm, resulting in incorrect results[13]. This could result in faulty assumptions, diagnosis, and therapies for patients.

III.OPEN PROBLEMS IN EXISTING SYSTEM

Although the outcomes of prediction are promising, these traditional methodologies are still far from being highly precise and efficient.The current mechanisms are simple and effective, but they are exceedingly sensitive to disruption[14]. Furthermore, cutting-edge approaches use only one algorithm, resulting in erroneous results. This could lead to physicians making incorrect assumptions and providing patients with incorrect diagnoses and treatments[15].Prediction systems for liver illness have played an essential role in people's lives, and many scholars believe it is an important topic[16]. Although the forecast findings are encouraging, these old methods are still far from being highly exact and efficient. Existing methods are simple and functional, but they are extremely vulnerable to interruption.Furthermore, cutting-edge approaches only use one algorithm, resulting in incorrect results[17]. This could lead to incorrect assumptions, diagnosis, and therapies for patients[18].

IV.PROPOSED SYSTEM

Machine learning is understandably one of the most widely used paradigms of big data management, where a significantly large set of distinct raw data can be effectively collated to make appropriate inferences and eventually to produce a typical collection of contextually useful collection of integrative information[19]. With the advent of the exponential technology expansion in the field of medicine, there is a perceived need to manage and utilise a massive quantity of data in order to develop effective and relevant inferences for doctors and patients.

V.ADVANTAGES OF THE PROPOSED SYSTEM

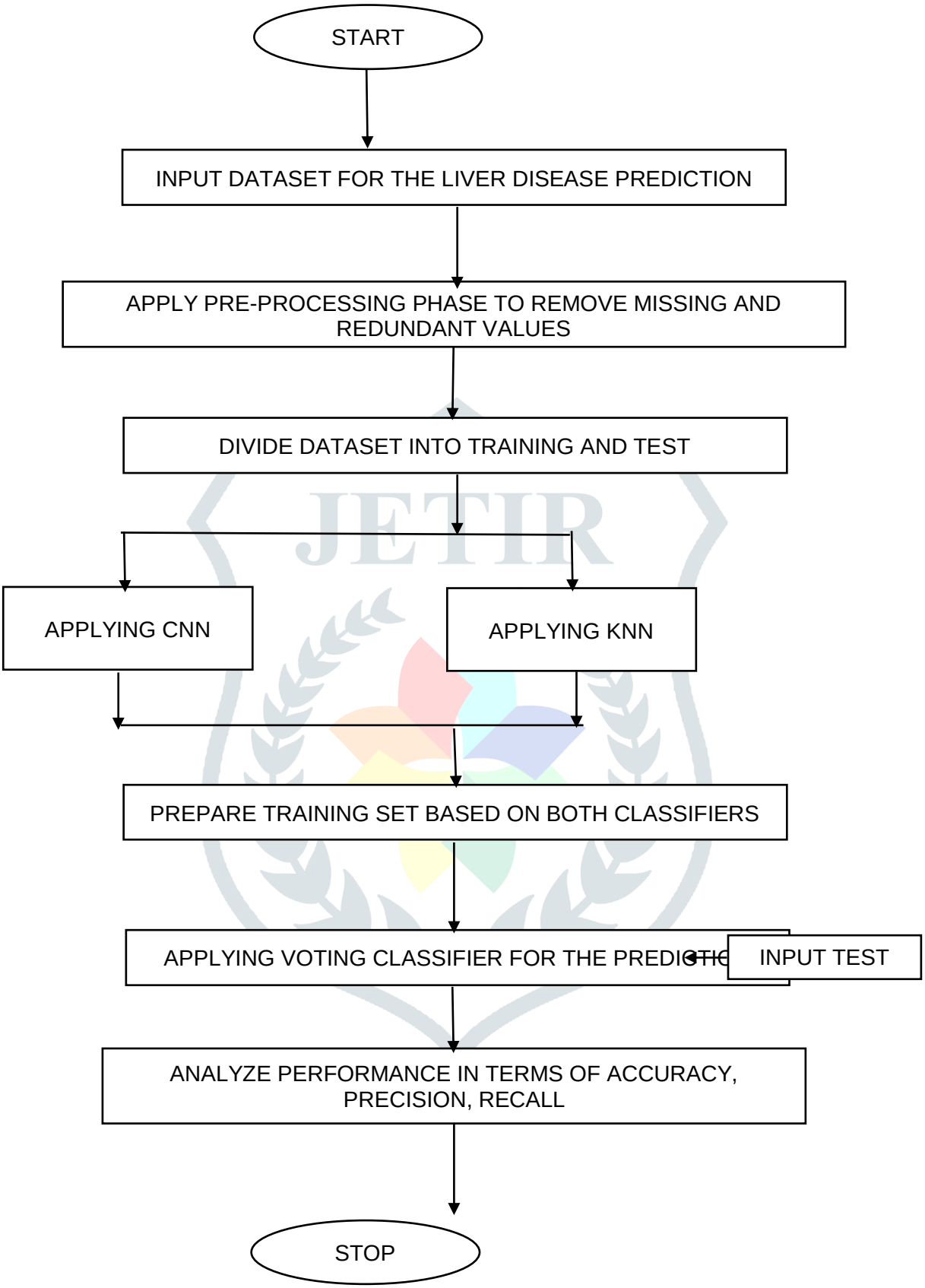
Considering the various differences that have been implemented in the current system, the following different advantages have been observed.The performance classification of liver-based diseases has been improved further: with a better understanding of the various types of ailments in the field of medicine, determining the type of liver disease and its occurrence has become a far less complicated task[20]. With developments in data mining paradigms and software architectures such as Hive and R facilitating the data collection process, the preprocessing and assessment steps are receiving more attention[21]. Various machine learning models can assess time complexity and accuracy, allowing us to measure different metrics based on the demands of the user: Every prediction system is built around the parameters that it is intended to receive, compare, and then use to make a prediction. As a result, multiple algorithms are utilised to model the predictive system depending on the circumstance[22]. The various machine learning algorithms determine the type of disease and the testing settings[23].

Various machine learning techniques with great accuracy of results: In comparison to other approaches discussed, the correct machine learning algorithm can effectively boost the efficiency of the predicted results. Machine learning models can anticipate risk variables early on: Machine learning algorithms forecast risk variables by analysing irregularities in the aggregate training data set and their related parameters using basic techniques.

VI.ADVANTAGES OF PROPOSED METHODOLOGY

A person's risk factor can be predicted using this process, and then the individual can receive therapy for their disease. Treatment is required since the occurrence rate of cardiovascular ailments is increasing at an unanticipated rate, and many people are unaware that the ML model will be more effective in inducing in the task. The ML model will be more effective at causing the task to be completed. As a result, the prevalence of cardiovascular disorders is expected to rise further.

FLOW CHART



The Flowchart Explains The Research Methodology

VII.THE SYSTEM HAS FOLLOWING ADVANTAGES

There is no need for medical knowledge: You do not need any prior understanding of medical science or liver illnesses to use this programme to anticipate liver disease. All you have to do is enter the details requested, which are already included in the blood test report (some, such as age and gender, are already known), and you will receive the prediction results.High accuracy: For the dataset that we used to create this application, the algorithm predicts the results with 100% accuracy. While the accuracy may differ in some circumstances, it will be high enough to be reliable on a broad scale.Immediate

results: The outcomes are anticipated here within seconds of entering the information.You do not have to wait for a doctor.

VIII.ADVANTAGES OF PROPOSED METHODOLOGY

A person's risk factor can be predicted using this process, and then the individual can receive therapy for their disease. Treatment is required since the occurrence rate of cardiovascular ailments is increasing at an unanticipated rate, and many people are unaware that the ML model will be more effective in inducing in the task. The ML model will be more effective at causing the task to be completed. As a result, the prevalence of cardiovascular disorders is expected to rise further.

IX.KNN

The KNN algorithm's implementation specifics are discussed in this section. The full training dataset serves as the KNN model. The KNN algorithm will look through the training dataset for the k-most comparable cases when a prediction is needed for an unobserved data instance. The most comparable examples' prediction attributes are compiled and sent back as the prediction for the unknown instance. The type of data has an impact on the similarity metric. The Euclidean distance can be applied to data with real values. The hamming distance method can be used to other forms of data, such as category or binary data. The instance-based, competitive learning, and slow learning algorithms family includes the KNN algorithm. For the purpose of making predictions, instance-based algorithms model the problem using data instances (or rows). All training observations are kept in the model through the KNN algorithm, which is an extreme instance-based technique. Its underlying use of competition amongst model components (data examples) in order to arrive at a prediction conclusion makes it a competitive learning algorithm. Every data instance competes to match or be the most similar to a certain unknown data instance and makes a forecast as a result of the objective similarity measure between data instances.

X.SVM

SVM looks for the best hyperplane to divide the data into distinct classes. SVM is implemented in Python using the scikit-learn library. The pre-processed data is divided into a training set and test set, each comprising 25% and 75% of the whole dataset. In a high- or infinite-dimensional space, a support vector machine creates one hyper plane or a collection of hyper planes. The hyper plane with the greatest distance from the closest training data point for any class (referred to as the functional margin) achieves a decent separation since, generally speaking, the higher the margin, the lower the classifier's generalisation error is.

XI.LOGISTIC REGRESSION

One of the more straightforward classification models is logistic regression. It is beneficial when experimenters wish to look at interactions between variables because of its parametric nature, which can be partially comprehended by looking at the parameters. A vector of parameters = (0, 1...) can fully define a parametric model. A straight line using the formula y = kx + m with the parameters k and m is an example of a parametric model. It is possible to reconstruct the entire model using known parameters. A parametric model called logistic regression has parameters that are coefficients to the predictor variables, expressed as 0 +1 +X1 +...PXp. where the intercept is designated as 0. Instead, we can denote X as the vector form of the aforementioned sum of the parameterized predictor variables. The name logistic regression is unfortunate because, unlike classification, where the response variable is discrete, regression models are typically utilised to find continuous response variables. The phrase may have been inspired by the fact that logistic regression revealed a continuous chance of the response variable falling into a certain class.

ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

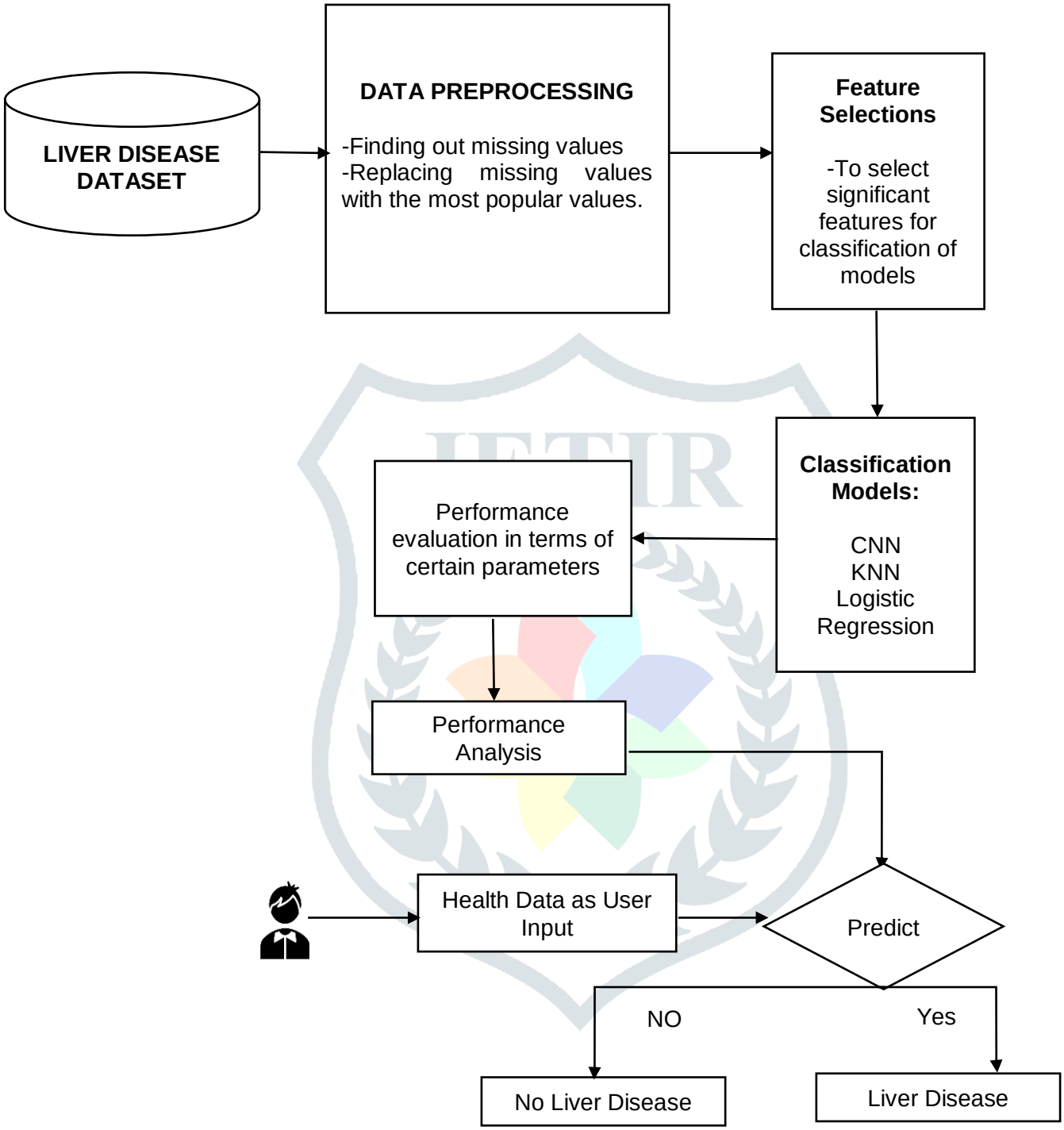


Figure 1: System Architecture

XII.RISK ANALYSIS OF THE PROJECT

FEASIBILITY STUDY

In this stage, the project's viability is determined by the increase in server performance, and a business proposal is presented with a very basic project design and some cost projections. The feasibility assessment of the suggested system must be completed during system analysis. Understanding the main system requirements is crucial for the feasibility analysis.

1. Economical feasibility

2. Technical feasibility

3. Operational feasibility

1.Economical feasibility

This study is being conducted to determine the system's potential financial impact on the organisation. The corporation is only able to invest a certain amount of money in the system's research and development. The expenses must be defended. Because the majority of the technologies were freely available, the produced system was also developed under the budget. Only the products that were customised needed to be bought.

2.Technical feasibility

This study is being done to evaluate the system's technical requirements, or technical feasibility. Any system created must not place a heavy burden on the technical resources at hand. As a result, the client will face high expectations. The created system must have reasonable requirements because its implementation only necessitates minor or no adjustments.

3.Operational feasibility

The goal of the study is to determine how much the user accepts the system. This includes the instruction needed for the user to operate the system effectively. The system shouldn't make the user feel threatened; instead, they should view it as a need. The techniques used to inform and acquaint the user with the system are the only factors that affect the level of acceptance by the users. As the system's ultimate user, his confidence must be increased so that he may offer some helpful criticism, which is encouraged.

XIII.THE SYSTEM HAS FOLLOWING ADVANTAGES

No medical knowledge is necessary: Using this application to forecast liver disease requires no prior medical science or understanding of liver problems. You may acquire the prediction results by just entering the information that is already in the blood test report (certain information, such age and gender, is already known).

High degree of accuracy: For the dataset we used to develop this application, the system predicts the outcomes with a 100% accuracy rate. Even though the accuracy might vary in some instances, it will still be sufficient to be relied upon on a wide scale.Results are projected in a matter of seconds after entering the information. There is no need to wait for a physician.

XIV.SOME MACHINE LEARNING METHODS

The two primary categories of machine learning methods are unsupervised and supervised. For input and the intended output, supervised algorithms require supervision by a person with machine learning expertise. The method will then be used on fresh data after the model training is finished. Unsupervised algorithms don't need any data training. They do, nonetheless, employ an iterative process. Deep Learning is the name of this strategy.

Artificial neural network is another name for unsupervised learning techniques. These networks are employed in situations where there is greater complexity because they are more versatile than supervised learning systems. Such neural networks advance by sifting through the training set and automatically identifying associations between both the dataset's parameters. Supervised machine learning algorithms are applied to the previously studied data in the past and then to new data. Such a system is able to provide outputs for any new input once sufficient training is done. This algorithm also compares its output with the correct, intended output and finds discrepancies, and then modifies the model accordingly.

On the other hand, unsupervised machine learning algorithms have been when the information has neither been classified nor labeled. This category studies how systems can make inferences into a function that describes a hidden structure in unlabeled data. But there is a drawback, this system doesn’t tell the right output, rather it explores the data and draws inferences from datasets.

Semi-supervised machine learning algorithms lie in between the above-mentioned systems. This is because they use both labeled and unlabeled data for training purposes –a smaller amount of labeled data and a larger amount of unlabeled data in combination. This system uses techniques that are able to improve accuracy. Generally, semi-supervised learning is opted when the acquired dataset (labeled) requires skilled resources to train it.

A learning method that engages with the surroundings is reinforcement machine learning. Generating activities, then looking for inconsistencies, is how it is done. Using this technique, behavioral patterns in the dataset can be contextually computed by machines and software agents to improve performance. The best way to proceed is then determined by analyzing simple reward responses. The aforementioned method is referred to as a reinforced signal.

Although it produces results more quickly and accurately, it may also require more time and resources during training. It is considerably more effective at processing massive datasets when deep learning, AI, and intelligent systems are combined.

XV.WORKING OF MACHINE LEARNING ALGORITHMS

The machine learning component is considered to be the brain of the system where all the learning aspects take place and are controlled centrally. The machine learning algorithms enable the system to learn, similar to how the human brain does. Human brains are used to understanding and making viable inferences using experiences. However, in order for a machine to make an accurate prediction, the following data could be utilized. The core activity phases of a machine learning system would be - learning and inference. The discovery of patterns plays a major role. Feature selection would be the follow-up procedure, where it is decided which of the core values of the field are put to use. The discovery part is facilitated with the collection of data, which is put to use. The right set of data is also critical at the feature selection stage. The list of these attributes is chosen by what is known as an attribute vector.

XVI.LITERATURE SURVEY

Anu Sebastian, Surekha Mariam Varghese, "Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver"Survival Analysis is an extensively used procedure in the field of medical science. The idea of being able to predict the life expectancy of the subject is of immense value and utility to both, the doctors and the patients. There are three preliminary steps that serve as the elementary foundation of any medical treatment paradigm. The diagnosis stage, the classification stage, the assessment stage, the conclusion stage and finally the treatment stage. All these stages are expected to be accurate to the parameters and effective in their measure to distinctly reflect the quantified magnitude and the intensity of the study of the disease in the context. One of the most widely used classification methodologies that have been used for an extensive assessment of liver diseases, particularly cirrhosis is the Child-Pugh classification method. It is understandable from the extensive study of a voluminous set of cases that the life expectancy of different patients, suffering from different intensities and kinds of liver cirrhosis, is different. Fuzzy Logic, for instance, suits the context like a tailor-made technique.
Insha Arshad, Chiranjit Dutta, "Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques"Liquor is expended in overabundance by a large number of individuals over the world. Liquor utilization is legitimately connected to perilous liver maladies, for example, cirrhosis which may at last lead to death. Early location of liver illness brought about by overutilization of liquor would help in sparring existences of numerous individuals. By distinguishing liver ailment in its beginning time, it very well may be analyzed in time and may prompt full recuperation in certain patients. This paper proposes identification just as to foresee the nearness of liver sickness utilizing information mining calculations. We will settle on a choice tree for the dataset and afterwards the principles will be created. Subsequent to deciding the principles, we will utilize diverse information mining calculations to prepare and test the dataset to distinguish the liver sickness. The information was gathered from UCI storehouse and our preparation dataset was created. It comprises of 7 unique qualities having 345 occurrences. In the dataset, distinctive classes of blood tests are taken into contemplations which are straight forwardly connected to liver illnesses that may emerge because of unnecessary liquor utilization alongside recurrence of liquor utilization. In light of the sort of liver sickness recognized, the forecast might be proposed.

XVII.INFERENCES FROM LITREATURE SURVEY

There are some logically strong inferences that can be made from the literature review. Since the thesis is to composite the ideology of using machine learning algorithms for the prognosis, diagnosis and study of liver diseases and their predictability, it is important to deal majorly with the kind of machine learning algorithms that would suit the purpose and be centric on the major objectives - being able to predict the presence of a liver disease in the most accurate possible way. The literature surveys conclude the use of Naive Bayes and Support Vector Machine algorithms for the prediction of liver diseases. There are two major parameters that are involved in understanding the suitability of the respective methodologies and they are - the time taken to execute the prediction process and the accuracy of the predictive result. It is clear through various studies and experimentations that SVM classifier is the best of all the algorithms owing to the extremely high accuracy rates. But when it comes to the time taken to execute the predictive process, the Naive Bayes classifier reflects higher suitability since it takes the least possible time to execute the process.
From the above-mentioned literature works, it is clear that there has been effective research on this topic has been done and many models have been proposed.
It is evident that the above-mentioned systems have their own pros and cons.
While some of the recent works involve hybrid technologies and provide better accuracies, they are still far from what is needed.
With higher accuracy, comes the need for low computational costs, high processing speed, and most of all, the convenience of use.

XVIII.IMPLEMENTATION

The current system employs various approaches to reach a significantly less accurate conclusion while maintaining the same purpose. The accuracy of the outcomes generated determines the qualitative superiority of various strategies over one another. To parametrically get a firm conclusion on the prediction of liver illness, many aspects of the data are used. For the classification of patients with liver cirrhosis, fuzzy logic has been established. The Child-Pugh score is used in gastroenterology to evaluate the prognosis of chronic liver disease, particularly cirrhosis. It was initially developed to foretell surgical mortality. Today, it is utilised to assess the prognosis, the strength of the recommended therapy, and the requirement for liver transplantation. Some utilise a modified version of the Child-Pugh score that takes into account the fact that these disorders are characterised by high levels of conjugated bilirubin. The highest limit for a single point is 68 mol/L (4 mg/dL), and the upper limit for two points is 170 mol/L (10 mg/dL). With the use of a comparative degree, the systems were created in a similar way. The standardisation of prefixed conditions is prefixed.
These conditions serve as benchmarks against which the training data sets are analysed, and inferences are drawn in accordance with the findings. The conclusions can be deduced using pure mathematical modelling under some practically plausible conditions. The predefined contextual circumstance, the occurrence of liver disease, can be predicted using the Bayes theorem for conditional probability. This style of approach has two important purposes. Creating a template that blatantly indicates the presence of the illness component in the test cases' report is the first step. The patterns found in the physiological conditions seen in various test scenarios are used to define the template. In order to reach a specific conclusion on the traits of patients of all types who have liver problems, data mining techniques are applied. In order to reach a specific

conclusion on the traits of patients of all types who have liver problems, data mining techniques are applied. The data pieces of patients who have experienced excessive drunkenness, which may have more likely been the cause of the sickness, would also be included in the data mining approaches.

XIX.REFERENCE

[1] A book reference of predicting the existence of diseases such as locomotor disorders
https://www.sciencedirect.com/science/article/abs/pii/S0049017222000658
[2] A website reference of demonstrating its potential for disruptive innovation
https://www.tandfonline.com/doi/abs/10.1080/20421338.2021.1975355
[3] An article reference of medical procedure working
https://www.mdpi.com/1660-4601/19/3/1122
[4] A book reference of relevant sets of useful medical information
https://onlinelibrary.wiley.com/doi/abs/10.1002/bdr2.2151
[5] A website reference of research initiatives in the same environment
https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-022-04447-4
[6] An article reference of aid in the overall process
https://www.sciencedirect.com/science/article/abs/pii/S221478532105570X
[7] A book reference of medical diagnosis and disease prediction.
https://link.springer.com/chapter/10.1007/978-981-19-5868-7_35
[8] A website reference of the liver is responsible for several vital tasks
https://zoologicalletters.biomedcentral.com/articles/10.1186/s40851-022-00200-7
[9] An article reference of energy and nutrition
https://bmcnutr.biomedcentral.com/articles/10.1186/s40795-022-006
[10] A book reference of enzyme levels in the blood
https://link.springer.com/chapter/10.1007/978-3-031-17774-3_10
[11] A website reference of methods is utilised to predict liver illness
https://onlinelibrary.wiley.com/doi/abs/10.1111/eci.13895
[12] An article reference of extremely vulnerable to interruption
https://www.sciencedirect.com/science/article/pii/S0951832022006858
[13] A book reference of algorithm resulting in incorrect results
https://www.sciencedirect.com/science/article/pii/S2405959522000923
[14] A website reference of exceedingly sensitive to disruption
https://link.springer.com/chapter/10.1007/978-3-031-19656-0_1
[15] An article reference of physicians making incorrect assumptions
https://www.sciencedirect.com/science/article/abs/pii/S0012369223000363
[16] A book reference of Prediction systems for liver illness
https://www.mdpi.com/2077-0383/12/8/3006
[17] A website reference of cutting-edge approaches only use one algorithm
https://www.tandfonline.com/doi/abs/10.1080/03772063.2022.2163928
[18] An article reference of diagnosis and therapies for patients
https://www.sciencedirect.com/science/article/pii/S1569199322000315
[19] A book reference of collection of integrative information
https://informationr.net/ir/27-1/paper922.html
[20] A website reference of performance classification of liver-based diseases
https://www.sciencedirect.com/science/article/pii/S1361841522003085
[21] An article reference of assessment steps are receiving more attention
https://www.tandfonline.com/doi/full/10.1080/02687038.2022.2163462
[22] A book reference of predictive system depending on the circumstance
https://www.sciencedirect.com/science/article/abs/pii/S0957417422016712
[23] A website reference of disease and the testing settings.
https://onlinelibrary.wiley.com/doi/full/10.1111/all.15571

BIBILOGRAPHY

Gudiwaka vijayalakshmi working as a Assistant professor in the department of Computer Science and Engineering sanketika vidya parishad engineering college, visakhapatanam Andhra pradesh. with 2 years of experience in Computer science and engineering (CSE) , accredited by NAAC. With her area of interests in java, python,.Net, HTML.

Rongala Rajesh is studying his 2nd year Master of Computer Applications in Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P.With her interest in Python, machine Learning and as a part of academic project she choose analysis and prediction of liver disease using CNN and KNN using Python. The article have been evolved from an idea to understand the flaws in conventional reporting and keeping time consistency, A full fledged project along with code has been submitted for Andhra University as an Academic Project.