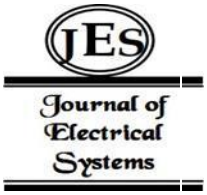
[1]J. Grace Arputha Rajakumari

[2]Dr. N. Balajiraja

An Efficient Decision-Making Disease Diagnosing System Using Healthcare Datamining Techniques



Abstract: - Healthcare uses Data Mining techniques for knowledge discovery and identifying successful prescription patterns for diseases and prediction using computer aided diagnosis or expert learning. Integrating Data Mining with forecasting can provide dependable and high quality forecasts. Prediction of diseases using data mining techniques is a motivating task for increasing diagnostic accuracy. Hence the objective of this research is in using data mining as they help decrease cost and time. Knowledge discovery from medical data is a complicated task, mainly due to irrelevant and unwanted data. Using more than one data mining technique for predicting diseases can also result in better accuracy. Hence the main objective of this research work is to predict Healthcare diseases from patient's records and suggesting a non-invasive data mining model. Moreover, Features provide state-of-the-art performance for recognition of abnormalities. While the accuracy of action recognition has been continuously improved over the recent years, the extraction of lesser number of features and subsequent identifications based on these extractions have been preventing methods from scaling up to real-life issues. This problem is addressed in this research work by the development of highly efficient features using feature information in disease recognitions. Moreover the speed of feature extraction and feature selection can help disease classification perform better at the cost of a negligible reduction in recognition accuracy. The main goal of this work is efficient disease recognition while exploring the speed-memory trade-off in feature extraction and selection.

Keywords: CART, Clustering, Diagnosis, Data mining, Decision Making, Healthcare.

I. INTRODUCTION

Vast amounts of data are generated during the healthcare process. While technological advances in the form of computer-based patient record software and personal computer hardware have made the collection and access of healthcare data more manageable, there are few tools available to enable ongoing evaluation and analysis of this clinical data once it has been captured and stored.

Evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance understanding of disease progression and management. Technology is needed to search for these patterns and relationships in large amounts of clinical data. Past efforts in this area have been limited primarily to epidemiological studies on data mining initiative and claims databases. Knowledge discovery in databases (KDD) is the search for global relationships and patterns that exist in large databases but are "hidden" in large amounts of data [1]. The typical data mining process involves transferring data originally collected in production systems to a data warehouse, cleaning or purifying the data to eliminate errors and check for consistency in format, and then searching the data using statistical queries, neural networks, or other machine learning methods [2]. Though many applications of KDD have focused on discovering novel data patterns to solve business related problems, they have also been used extensively in healthcare researches. data mining has been used is used to discover subtle factors affecting the success and failure of back surgeries and treatments which led to improvements in patient care [3].

MATLAB can work with matrices, deleting a row, a column, transposing a matrix, calculating the determinant…etc. Data mining systems have evaluated identifications and intervention strategies of diseases that were likely to cut costs. Thus, the eventual goal of knowledge discovery effort is to identify factors that can improve the quality and reduce costs in mining healthcare information. Hence, this research work uses data mining techniques which are tested with MATLAB.

[1]*Research Scholar J.J college of Arts and Science (Autonomous) Affiliated to by Bharathidasan university, Tiruchirapalli* Pudukkottai,India

rajigrace81@gmail.com

[2]*Assistant Professor PG and Research Department of Computer Science J.J college of Arts and Science(Autonomous) Affiliated to by Bharathidasan university, Tiruchirapalli* Pudukkottai,India

nbalajiraja@gmail.com

II. LITERATURE SURVEY

Abdelhamid A *et al.,* (2023) designed a system to diagnose hepatitis disease by using genetic neural network methodology. The clinical symptoms were used by the researchers as input in the developed diagnostic system. The performance of the designed diagnostic system to diagnose the Hepatitis B disease has been measured on the basis of classification accuracy, and the observed accuracy is 99.14%.

Admass, W *et al.,* (2022) compared the disease-based approach in Data mining with database reasoning and used the data mining technique to build a decision support system based on evidence to minimize the unnecessary testing to reduce the total expense of patient care using healthcare techniques.

Al-Dafas, M *et al.,* (2022) used the term frequency-inverse document frequency approach for developing a strategy for reducing the support for sensitive item sets. A transaction containing a large number of sensitive items but having minimal influence on other transactions is very likely to be updated. The deletion priority is determined by the number of sensitive items supported.

Alhasani, A *et al.,* (2023) proposed a fully automated computer-aided Healthcare diagnosis system to classify malignant and benign masses using breast magnetic resonance imaging. The texture features were selected by integration of support vector machine with Relief feature selection method. This system achieved an accuracy of 92.3%.

Chengathir, S *et al.,* (2022) use the technique of Healthcare data mining for supporting the users to know the answers for predefined questions in the application of web based. For diagnosing heart disease diagnosing system doctors use these intelligent decisions when NB algorithm accuracy can be improved by utilizing various techniques Decision techniques for heart disease prediction.

Deb, A., *et al.,* (2022) introduced the Internet of Things and machine learning in the healthcare domain for the diagnosis Decisions disease. The proposed framework was primarily focused on assessing recent memory loss in human conversations. The simulations showed that the proposed framework was highly efficient for diagnosing and predicting life-threatening diseases.

Dutta, K *et al.,* (2022) suggested an data mining method for predicting cardiac disease. The suggested research makes use of the Cleveland heart disease dataset, as well as data mining methods including classification and regression. RF and DT machine learning algorithms are used the disease diagnosing model's innovative method was developed.

Fitriyani, N. L *et al.,* (2024) introduced a system with an healthcare optimization method to eliminate privacy concerns and enhance the detection technique for heart disease identification. Using an updated federated Techniques learning method for user sites and the cloud, they create and suggest a privacy-aware system for predicting heart disease in healthcare.

Hajjej, F *et al.,* (2024) developed Cardio a method that recognizes the existence of healthcare in a patient predictive models for heart disease patients, indicating the need for combinational and more complex systems to improve the accuracy of identifying the early onset of heart disease.

III. PROBLEM STATEMENT

Human population explosion has resulted in the manifestation of many new and unknown diseases, where certain diseases do not have a permanent cure. Treating Healthcare diseases are a major challenge to clinicians as the Signs and symptoms of disease overlap with other disorders. Some disease results in severe complications of other body parts like nephritis, vacuities, pulmonary hypertension, interstitial lung disease and strokes [4], thus making diagnosis complex A patient's feelings of distress, guilt, and anxiety occur in their negative social experiences when they fail to understand the intensity of their illness [5]. Moreover, patients experience unique challenges personally, when diagnosed with an invisible illness that threatens their quality of life [6]. To improve a patients' emotional and physical health, an awareness on Healthcare diseases in patients and directions in diagnosis of disease are needed. Patients at an increased risk in their psychiatric conditions and decreased Quality of Living, add to the issues. Patients have to quickly identify and take care of exhibited physical symptoms for managing disease effectively. Thus, though disease can be treated effectively, its

diagnosis early is a major problem and mainly due to its manifestations as other diseases. In processing data for identifying disease, machine learning techniques need previous history of patients in large numbers for their training which is an issue as the availability of such data for research is scarce. Though current improvements in Medicare have enhanced patient's survival rates, fear of mortality remains a major issue as patients suspect treatment methods. Hence, disease with limited historic data, its complications and awareness are major issues for diagnosing in patients early.

IV. PROPOSED WORK

- To extend the life span of a chronically diseased patient.
- Automatically predict patient outcomes based on recorded symptoms.
- Propose an intelligent clinical decision-making model to reduce errors in disease decision-making.
- To design and propose new algorithms and techniques to predict diseases early.
- To study and analyze lupus patients real-time data set and turn it into useful knowledge.
- To increase the speed of predictions while reducing the cost of automated predictions by selecting fewer and important features
- To analyze existing algorithms and compare the proposed method's effectiveness.

V. METHODOLOGY

5.1 Dataset

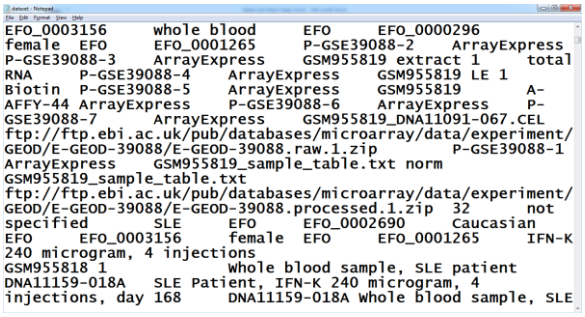


Fig 5.1: Sample Dataset

5.2 Data Cleaning

Data cleaning is an essential part of machine learning as it plays an important role in building models. Data cleaning can make or break an analysis. Professional data analysts spend a lot of time in this step. A clean dataset can get desired results even with a simple algorithm, which is beneficial. Figure 5.2 depicts the flow of Data Cleaning.

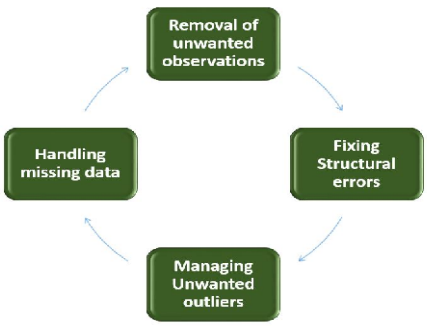


Fig 5.2: Data Cleaning Flow

Data cleaning involves different steps for different data. The steps followed in this research work is Missing value prediction, Redundancy avoidance, Filtering (Fill mean mode value) and Attribute reduction.

5.3 Missing Value Prediction

Missing data can be identified in three ways as detailed below

- **Missing Completely At Random (MCAR):** Random missing values are the highest level of randomness where features are not dependent on any other features values.
- **Missing At Random (MAR)**: Missing values in features that depend on the values of other features.
- **Missing Not at Random (MNAR):** This means that the data collection process must be validated.

5.4 Dataset Preparation

A dataset is a collection of patient data. A dataset corresponds to the contents of a single database table or a single statistical data matrix, where each column of the table represents a specific variable and each row corresponds to a specific member of the dataset. Machine Learning methods use a training data set where actual data is used to train the proposed model for performing various actions. The training data set applies concepts like neural networks for learning and expected results. It includes both input and expected output data. Training sets make up the majority of the total data. In this work the training constitutes 70%. The testing model adapts to fit to parameters in a process called adjusting weights. The test data set is then used to evaluate how well a machine learning technique was trained with the training data set..

5.5 Decision Making

Making the right decision is often a challenge. A simple and quick approach for taking a decision is following past experiences in similar situations. The human brain decides based on two factors namely logical and intuitive. Most decisions are automatic responses because the logical part invents the reasons for the decision. The intuitive system also decides on thousands of decisions, but may be biased toward the last unsuccessful outcome. Tools like decision matrix can help in unbiased decision making. It is an advanced approach for making decision and scores each possible option against some criteria or feature. This approach results in creating decision matrix for analysis of possible options. Machine learning techniques help improve decision making and can be viewed as assigning or predicting correct label based on data features (Classification Problem).

VI. EXPERIMENTAL RESULTS

6.1 Data Cleaning

The fields taken and first analyzed for Missing values. Missing values can change the course and direction of a result, if not handled. Hence, first step is handling missing value. Figure 6.3 depicts the output of Missing Values

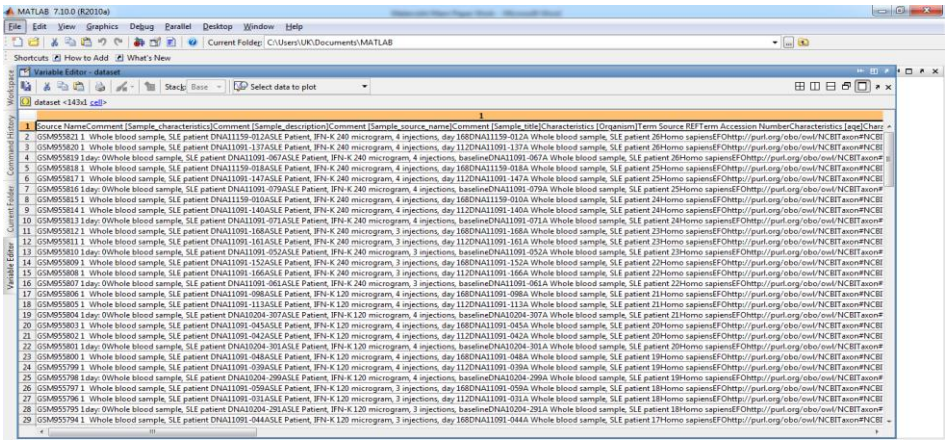


Fig 6.1: Data Cleaning Results

6.2 Decision Making

Decision trees are one of the best-known decision-making techniques, probably because of their inherent ease in visually communicating an option or set of options and their associated uncertainties and outcomes. Their simple structure makes them useful for a wide range of applications. They can be drawn by hand to help quickly outline and communicate the key elements of a decision. Alternatively, the simple logical structure of a decision tree allows it to solve complex multiple decision scenarios and problems with the help of computers. The proposed work uses the CART algorithm to form a decision tree based on the criteria listed in Table 6.1.

Table 6.1 Decision Making Criteria

Attribute	Decision Weight
General Attributes	1
Disease Activity	2
Symptoms	3
Test Results	4

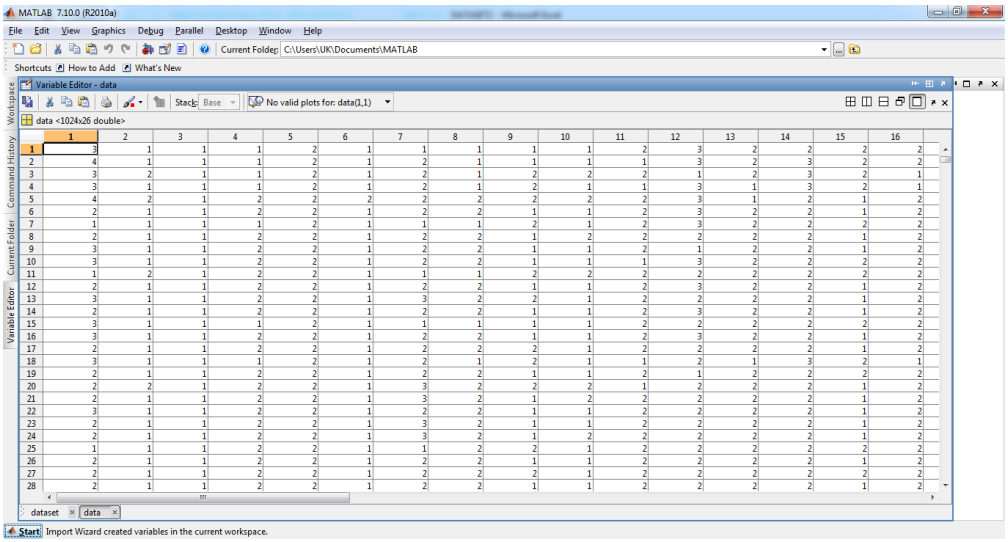


Fig 6.2: Decision Making Criteria Results

6.3 Proposed HKCART Algorithm

Input: The Healthcare dataset with n entities.

Output: A set of optimal clusters K_i.

Step 1: Likewise any abnormal or out of range values are also not considered for preprocessing.

Step 2: Records of incorrect and missing data are not considered and assign mean mode value.

Step 3: If Multi variant attributes or more than one instance are there then

Step 4: Remove redundant value using deletion query operation

Step 5: Normalization of missing attribute instances is done by filtering.

Step 6: Processed, filtered value converted as MAT file and stored in separate database.

Step 7: Initialize dataset $\sum (F) = \{ f_1, f_2, f_3 \ldots f_n \}$ attributes.

Step 8: Identify the Outliers in the considered column $\sum (F') = \{ f_1', f_2', f_3' \ldots f_n' \}$

Step 9: Repeat, formulate the rules for identifying the similar attributes.

Step 10: do until, Identify the frequent itemsets.

Step 11: Specify the threshold Mean, proportion value.

Step 12: Identify the K initial mean vector from the attributes

Step 13: Identify the distance between f_i attributes and the centroid value f_j.

Step 14: Recalculate until new centroid f_j identified.

Step 15: Identify the end convergence.

Step 16: Fin the neighborhood active attribute rule set.

Step 17: Generate recommendations from most frequent itemset.

Step 18: Identify the disease threshold mean prediction value.

Table 6.2 Comparative performances of proposed work

Algorithms	**Sensitivity**	**Specificity**	**Accuracy**
CART	97.33	97.66	97.19
K-Means	93.23	93.13	93.56
Proposed HKCART	98.33	98.66	98.45

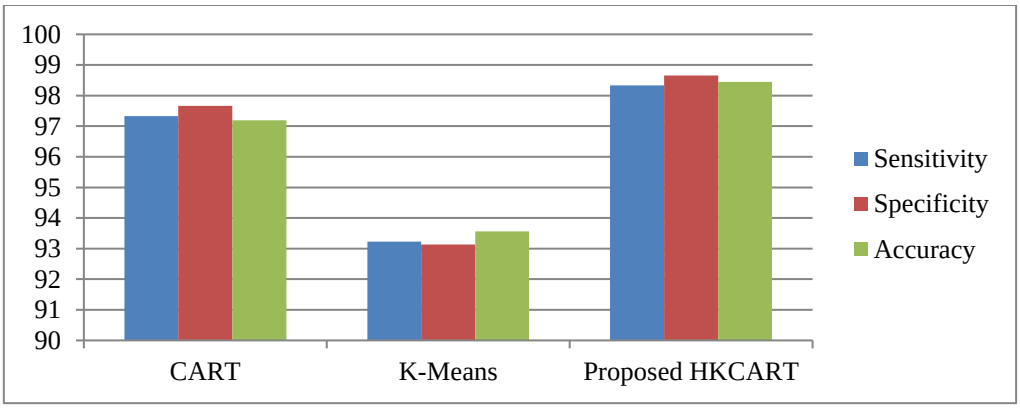


Fig: 6.3 Performance evaluations of proposed work

VII. CONCLUSION

The work has proposed and demonstrated the need for innovative technique to detect diseases from test results. Clustering though an easy and usable technique has its own demerits in disease predictions. This paper has demonstrated ways of overcoming those limitations with decision trees and rule based techniques which have helped improve efficiency and accuracy of disease predictions. The proposed techniques are implementable in computer aided systems. Moreover, the proposed techniques attempt to eliminate deficiencies in automatic predictions of Healthcare diseases. In the future, the technology could be expanded to mine predictive data for other diseases, as long as doctors identify criteria or parameters in the test results. Thus this work concludes that the proposed techniques can predict complicated diseases in a noninvasive way and from clinical lab test results.

References

[1] Abdelhamid, A. A., Eid, M. M., Abotaleb, M., & Towfek, S. K. (2023). Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques. *Journal of Artificial Intelligence and Metaheuristics*, *4*(2), 45-53.

[2] Admass, W. S. (2022). Developing knowledge-based system for the diagnosis and treatment of mango pests using data mining techniques. *International Journal of Information Technology*, *14*(3), 1495-1504.

[3] Deb, A., Koli, M. S. A., Akter, S. B., & Chowdhury, A. A. (2022, June). An outcome based analysis on heart disease prediction using machine learning algorithms and data mining approaches. In *2022 IEEE World AI IoT Congress (AIIoT)* (pp. 01-07). IEEE.

[4] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2024). HDPM: an effective heart disease prediction model for a clinical decision support system. *IEEE Access*, *8*, 133034-133050.

[5] Hajjej, F., Ayouni, S., Alohali, M. A., & Maddeh, M. (2024). Novel framework for autism spectrum disorder identification and tailored education with effective data mining and ensemble learning techniques. *IEEE Access*.

[6] Harouni, M., Karimi, M., Nasr, A., Mahmoudi, H., & Arab Najafabadi, Z. (2022). Health monitoring methods in heart diseases based on data mining approach: A directional review. In *Prognostic models in healthcare: Ai and statistical approaches* (pp. 115-159). Singapore: Springer Nature Singapore.

[7] Jyothi, V. K., & Sarma, G. R. K. (2022, December). A combinatorial approach: Datamining and an efficient Deep neural network for Heart disease prediction. In *International Conference on Intelligent Systems Design and Applications* (pp. 533-542). Cham: Springer Nature Switzerland.

[8] Mohamed, R. R., Nasr, M. M., & ElSeddawy, A. I. (2023). Enhanced Detection of Heart Diseases Using Data Mining Techniques.

[9] Alhasani, A. T., Alkattan, H., Subhi, A. A., El-Kenawy, E. S. M., & Eid, M. M. (2023). A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining. *Methods*, *7*(9).

[10] Munaye, Y. Y, & Admass, W. S. (2024). Integrating case-based and rule-based reasoning for diagnosis and treatment of mango disease using data mining techniques. *International Journal of Information Technology*, *16*(3), 1699-1715.

[11] Naresh, V. S., & Thamarai, M. (2023). Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1490.

[12] Saeeda, M. G., & Jasimb, Y. A. Developing a Software for Diagnosing Heart Disease via Data Mining Techniques.

[13] [20] Santos-Pereira, J., Gruenwald, L., & Bernardino, J. (2022). Top data mining tools for the healthcare industry. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 4968-4982.

[14] Al-Dafas, M., Albujeer, A., Hussien, S. A., & Ibrahim, R. K. (2022). On the adaption of data mining technology to categorize cancer diseases. *Int J Artif Intell Inform*, *3*(2), 80-91.

[15] Jayasri, N. P., & Aruna, R. (2022). Big data analytics in health care by data mining and classification techniques. *ICT Express*, *8*(2), 250-257.

[16] Suresh, T., Assegie, T. A., Rajkumar, S., & Kumar, N. K. (2022). A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model. *Int J Elec Comp Eng*, *12*(2), 1831-1838.

[17] Yadav, D. C., & Pal, S. (2022). Thyroid prediction using ensemble data mining techniques. *International Journal of Information Technology*, *14*(3), 1273-1283.

[18] Zeinalnezhad, M., & Shishehchi, S. (2024). An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients. *Healthcare Analytics*, *5*, 100292.

[19] Chengathir, S. M., Bhuvaneswari, T., Maruthupandi, J., & Naga, P. R. (2022). Prediction of dengue using data mining classification algorithms. *International Journal of Health Sciences*, (I), 11860-11871.

[20] Nayim, M. A. M., Alam, F., Rasel, M., Shahriar, R., & Nandi, D. (2022). Comparative Analysis of Data Mining Techniques to Predict Cardiovascular Disease. *International Journal of Information Technology and Computer Science*, *14*(6), 23-32.

[21] Dutta, K., Chandra, S., & Gourisaria, M. K. (2022). Early-stage coronary ailment prediction using dimensionality reduction and data mining techniques. In *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021* (pp. 721-733). Springer Singapore

[22] Khafaga, D. S., Ibrahim, A., Towfek, S. K., & Khodadadi, N. (2023). Data Mining Techniques in Predictive Medicine: An Application in hemodynamic prediction for abdominal aortic aneurysm disease. *Journal of Artificial Intelligence and Metaheuristics*, *5*(1), 29-37.

[23] Safdari, R., Deghatipour, A., Gholamzadeh, M., & Maghooli, K. (2022). Applying data mining techniques to classify patients with suspected hepatitis C virus infection. *Intelligent Medicine*, *2*(04), 193-198.

[24] Shanshool, A. M., Saeed, E. M. H., & Khaleel, H. H. (2023). Comparison of various data mining methods for early diagnosis of human cardiology. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *12*(3), 1343-1351.

[25] Yang, J., & Hussein Kadir, D. (2023). Data mining techniques in breast cancer diagnosis at the cellular–molecular level. *Journal of Cancer Research and Clinical Oncology*, *149*(14), 12605-12620.