Advancements In Disease Prediction And Diagnosis: Leveraging Linear Regression And Pattern Matching In Medical Informatics

N.S.Kavitha[1] ,B.Dhivya[2] ,M.Sakthivel[3] ,G Revathy[4*]

[1]AP/IT Dr.N.G.P. Institute of Technology, Coimbatore, Email: Nskavi17@gmail.com
[2]Assistant Professor, Department of Artificial Intelligence and Data Science, Karpaga Vinayaga College of Engineering and Technology, Kanchipuram ,Email: dhivyabaskar91@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Erode Sengunthar Engineering College Thudupathi,Perundurai, Email: sakthivelmcs@esec.ac.in
[4]Assistant Professor Department of CSE Srinivasa Ramunjan Centre SASTRA Deemed University Kumbakonam, Email:revathyjayabaskar@gmail.com
*Corresponding Author

Received: 08.04.2024 Revised : 12.05.2024 Accepted: 22.05.2024

ABSTRACT
A notable development in medical informatics is the creation of disease prediction systems that combine pattern matching with linear regression. This method offers a comprehensive framework for early illness detection and prognosis. It is broken down into several modules, including knowledge discovery in medical systems, symptom identification (e.g., headache, fever, body pain, backache, swelling of joints, and running nose), pattern matching, and differential diagnosis. Healthcare workers can anticipate and diagnose diseases more effectively by using this methodology, which makes use of past data to find patterns within symptom presentations. This strategy has the potential to transform illness care and enhance patient outcomes by combining a thorough understanding of medical symptoms and their presentations with machine learning techniques like linear regression.

Keywords: Healthcare, Machine Learning, Disease Prediction, Public Health, Medical Conditions

I. INTRODUCTION

The potential of disease prediction to revolutionize healthcare lies in its ability to enable early intervention, enhance patient outcomes, and alleviate the strain on healthcare systems [1]. This, in turn, will lead to a more proactive and healthier approach to wellness and healthcare management. The objective of this project is to examine the changing landscape of disease prediction, emphasizing the significance of data-driven approaches and their potential impact on public health. We will explore the fundamental methodologies, technological advancements, and ethical considerations related to disease prediction, and analyze how this emerging field is positioned to revolutionize healthcare. The emphasis will be on the importance of proactive prevention and personalized care. The healthcare industry is an essential component of modern society, comprising a diverse network of professionals, institutions, and technologies dedicated to promoting the well-being and treatment of individuals. It is a complex and constantly evolving ecosystem that addresses a wide range of medical needs, from preventative care and diagnosis to treatment and rehabilitation [2]. Healthcare not only plays a fundamental role in preserving human health and reducing suffering, but it is also a significant driver of scientific and technological advancement. In this overview, we will explore the intricate world of healthcare, examining its vital role in society, the obstacles it faces, and the innovations that are revolutionizing the way we access, provide, and experience healthcare services. As we navigate the complex terrain of healthcare, we will emphasize the importance of collaboration, technology, and patient-centered care in shaping the future of this critical sector.

Machine learning [3], a subfield of artificial intelligence, has emerged as a transformative force in our increasingly data-driven world. This discipline empowers computers to learn from vast amounts of data, enabling them to identify patterns, make predictions, and adapt to new information without explicit programming. The applications of machine learning are widespread, spanning various domains such as finance, healthcare, transportation, and entertainment [4]. This technology revolutionizes the way we tackle complex problems and automate decision-making processes. In this introduction, we will explore

the profound impact of machine learning, from its core principles to its practical applications in the real world. Additionally, we will delve into its role in shaping industries, enhancing our daily lives, and paving the way for a future characterized by intelligent, data-driven solutions [5]. Throughout this journey into the realm of machine learning, we will uncover its potential, confront its challenges, and embrace the exciting possibilities it offers for driving innovation and advancing our understanding of the world around us.

II. LITERATURE REVIEW

Shameem Fathima [6] et.al., has proposed in this paper in this paper, we present a performance analysis of various data mining techniques for predicting the Arboviral Disease-Dengue. The data set used for this analysis consists of real-time data collected from super specialty hospitals and diagnostic laboratories. These data were obtained from blood samples taken during diagnostic investigations at study enrolment and again at hospital discharge. The data set contains 5000 records with 29 parameters. Our investigation focuses on two data mining techniques: SVM and Naive Bayes Classifier. To identify a small set of parameters for diagnostic purposes in clinical practice, we employ a proficient methodology called random forest classifier, which utilizes Gini feature importance. This methodology allows us to obtain the smallest possible set of symptoms that can still achieve decent predictive performance for the dengue disease. By combining both approaches, we evaluate the performance of the classifiers. The comparison between the methods reveals that SVM outperforms Naïve Bayes in diagnosing Dengue disease. Data mining classification techniques have the potential to be useful for medical diagnosis decision support in a clinical setting. Chikungunya, a crippling disease caused by a virus belonging to the family Otariidae, is an arbovirus that shares the same vector as the dengue virus. Therefore, in dengue endemic regions, chikungunya is also a significant cause of viral fever outbreaks associated with severe morbidity.

This paper by USAM [7] et.al. proposes that dengue fever is a prevalent mosquito-borne disease worldwide, and detecting outbreaks can aid in efforts to prevent its rapid spread. To improve predictive accuracy of dengue outbreaks, this research identifies the best features using three different feature selection algorithms (PSO, GA, and RS) and applies three predictive modelling techniques (J48, DTNB, and Naive Bayes). The study uses a dataset from the Public Health Department in Malaysia and shows that applying feature selection before predictive modelling improves accuracy and identifies a set of features for dengue outbreak detection for Malaysian health agencies. Dengue fever is a serious and life-threatening disease with millions of cases worldwide each year. Numerous researchers have identified various factors that can be used to predict future dengue outbreaks. Early detection of dengue outbreaks has been prioritized in previous studies, achieving predictive accuracies ranging from 70% to 80% (Long et al. 2010; Mousavi et al. 2011). Previous research has attempted to model the problem by utilizing historical data on dengue transitions and employing different classification techniques to develop predictive models for dengue outbreaks. This study aims to utilize historical dengue data to capture the relationships and patterns within the data, thereby enhancing the accuracy of predicting dengue outbreaks. Unlike most previous studies, this research first determines the most relevant attributes of the dengue data through a feature selection process before modeling the problem.

In this paper, Sanjay Kumar [8] et.al. propose the need for clinical care by medical practitioners or professionals in healthcare for various liver disorders. The research aims to demonstrate the classification mode of a dataset containing liver disorder patients in order to accurately predict liver diagnosis using different data mining algorithms. The authors analyse a real patient dataset to construct classification models for forecasting liver diagnosis. Five classification algorithms are applied to the given dataset, and parameters such as precision, recall, and accuracy are analysed to evaluate the performance of these classifiers. The study investigates methods to improve the performance of the classification models, ultimately showing promising results in diagnosing liver disease at its early stages. The liver is one of the largest organs in the human body, weighing approximately 3 lbs.

In this paper, Gulia [9] et.al. propose the use of classification techniques in the medical field to achieve more accurate classification compared to individual classifiers. The focus of this paper is on Liver Patient Classification, where computational intelligence techniques are employed. The authors evaluate various classification algorithms, namely J-48, Multi-Layer Perceptron, Support Vector Machine, Random Forest, and Bayesian Network, using liver patient datasets. To improve the prediction accuracy of liver patients, the paper introduces a hybrid model construction and conducts a comparative analysis in three phases. In the first phase, the classification algorithms are applied to the original liver patient datasets obtained from the UCI repository. In the second phase, feature selection is utilized to obtain a subset of liver patient data that consists only of significant attributes. The selected classification algorithms are then applied to this subset.

Dr. S. Vijayarani [10]et.al has in recent years, the utilization of data mining in healthcare sectors has become increasingly prevalent for disease prediction. Data mining involves extracting valuable information from large datasets, warehouses, or other repositories. Predicting diseases from voluminous medical databases poses a significant challenge for researchers. To address this issue, researchers employ data mining techniques such as classification, clustering, and association rules. The primary objective of this study is to predict liver diseases using classification algorithms. Specifically, the Naïve Bayes and support vector machine (SVM) algorithms are utilized. These classifier algorithms are compared based on their performance factors, namely classification accuracy and execution time. The experimental results indicate that SVM outperforms Naïve Bayes as a classifier for predicting liver diseases. Researchers in the healthcare sector face increasingly complex tasks when it comes to disease prediction from extensive medical databases. Data mining has become an indispensable tool in this field.

III. RELATED WORK

The branch of medical analysis known as computer-aided diagnosis, or CAD, is rapidly developing and multifaceted. The development of computer-aided diagnostic applications has garnered significant attention in recent years due to the potential for seriously misleading medical therapies resulting from errors in medical diagnosis systems. A crucial component of computer-aided diagnostic testing is machine learning (ML). An easy equation cannot be used to appropriately identify objects, such as body organs. For this reason, learning from examples is basically necessary for pattern recognition. Pattern recognition and machine learning hold the potential to enhance the accuracy of disease approach and detection in the field of biomedicine. They also honor the impartiality of the decision-making process. Machine learning (ML) offers a credible method for creating better, automated algorithms for the analysis of high-dimensional, multi-modal biomedical data. This survey paper presents a comparative analysis of different machine learning algorithms for the diagnosis of different diseases, including diabetes and heart disease. It demands attention to the assortment of machine learning algorithms and strategies utilized in sickness diagnosis and decision-making procedures.

IV. MATERIAL AND METHODS

The suggested system is a thorough framework created to transform medical informatics' ability to forecast and diagnose diseases. It is made up of multiple interconnected modules, each of which adds to the overall efficacy of the system. The cornerstone is the Knowledge Discovery module, which uses data mining and analysis methods to find important patterns and insights in medical data. This module makes it easier to identify the important variables and patterns that are necessary for creating reliable illness prediction models by sorting through large datasets. Common symptoms linked to a range of disorders are found and examined in the Symptoms module. With the help of this module, medical professionals can alter symptom profiles, increasing the accuracy of illness prediction models. This module facilitates the detection of possible illnesses based on symptom presentations by establishing a correlation between observed symptoms and established disease patterns. Finding connections or patterns between known diseases and reported symptoms requires the use of pattern matching. Healthcare practitioners can improve the accuracy of disease identification by validating or modifying disease predictions by examining how symptoms correspond with established disease patterns. The Differential Diagnosis module helps medical practitioners determine the possibility of different diagnosis by assigning probabilities to suspected diseases based on symptom presentations. This module helps with prioritizing treatment options and directing additional diagnostic investigations by providing a numerical estimate of the likelihood of each possible diagnosis.

A. Knowledge Discovery in Medical Systems

The goal of this module is to find meaningful patterns and insights in medical data. Healthcare workers can obtain knowledge that helps with illness prediction and diagnosis by utilizing historical data and applying strategies like data mining and analysis. In medical systems, the first step in creating efficient illness prediction models is knowledge discovery.

B. Symptoms

This module analyzes and identifies a number of symptoms that are frequently connected to illnesses, including fever, headaches, and physical pain. Healthcare practitioners can adjust disease prediction models by adding, removing, or changing symptoms. This module makes it easier to identify possible illnesses based on symptom presentations by mapping symptoms to recognized disorders.

C. Pattern Matching

In order to find correlations or patterns, pattern matching entails comparing symptoms of recognized disorders with observed symptoms. Healthcare practitioners can validate or improve illness forecasts by examining how symptoms correspond with established disease patterns. This module improves the precision of disease detection by recognizing recurring patterns of symptoms linked to certain diseases.

D. Differential Diagnosis

Based on symptom presentations, the differential diagnosis module assigns probability to possible diseases. Medical practitioners use symptom patterns, past medical history, and other pertinent information to determine the likelihood of different diseases. This module helps with prioritizing treatment options and directing additional diagnostic investigations by providing a numerical estimate of the likelihood of each possible diagnosis.

V. RESULT ANALYSIS

The disease prediction system's effectiveness in ranking possible diagnoses according to symptom presentations is highlighted by the outcome analysis. Strong relationships with observed symptoms are evident in the significantly high odds of incidence of diseases like malaria and influenza. On the other hand, diseases such as acute lobar pneumonia and basilar dysentery have smaller probability, indicating a reduced chance of occurring. The modest probability linked to upper respiratory infections, acute gastroenteritis, and sinusitis provide important information for therapeutic decision-making. Healthcare practitioners can more effectively identify and prioritize probable illnesses with this systematic method to disease prediction, which makes use of linear regression and pattern matching techniques. This improves the accuracy and efficacy of disease treatment measures. The probability of different diseases based on symptom presentations are displayed in the disease prediction system's outcome analysis. With comparatively high probabilities of 90.91% and 73.33%, respectively, diseases like influenza and malaria show a high probability of occurring given the symptoms that have been seen. On the other hand, diseases such as acute lobar pneumonia and basillary dysentery had lesser odds of occurring, with 31.82% and 38.89%, respectively. Acute gastroenteritis, upper respiratory infections, and sinusitis all have mid-range probabilities, which suggest moderate likelihoods. These findings demonstrate how the system may rank possible diagnoses according to symptom patterns, helping medical professionals make well-informed choices about patient care and course of treatment.

Table 1. Comparison table

Disease	Probability
Malaria	73.33
Influenza	90.91
Sinusitis	57.14
Acutegastroentite	53.85
Upper respiratory	69.23
Migrane	44.44
Acutelobarpnemonia	31.82
Bascillarydysentry	38.89
Pulmonarytuberclosis	35
Hypertension	33.33

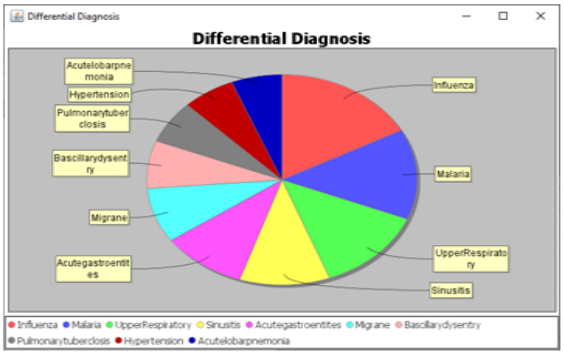

Figure 2. Comparison graph

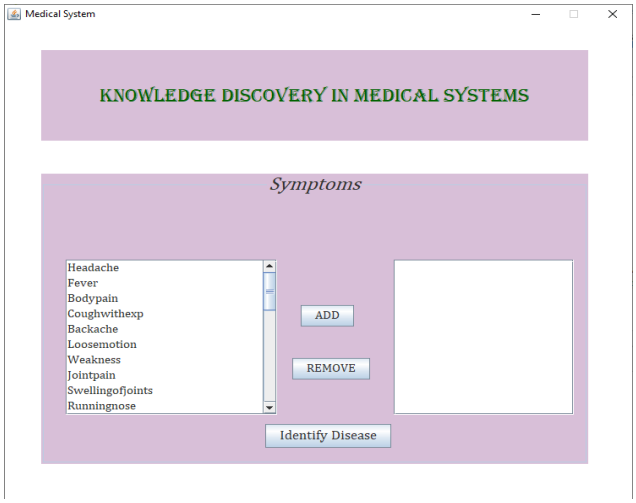

Figure 3. Medical Systems Figure

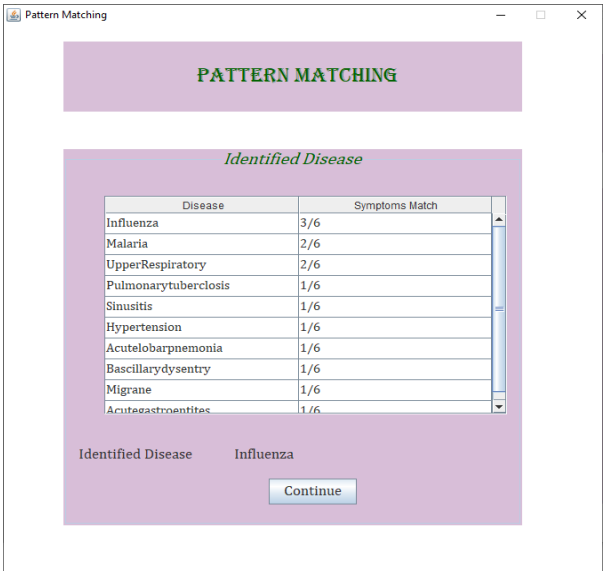

Figure 4. Pattern matching

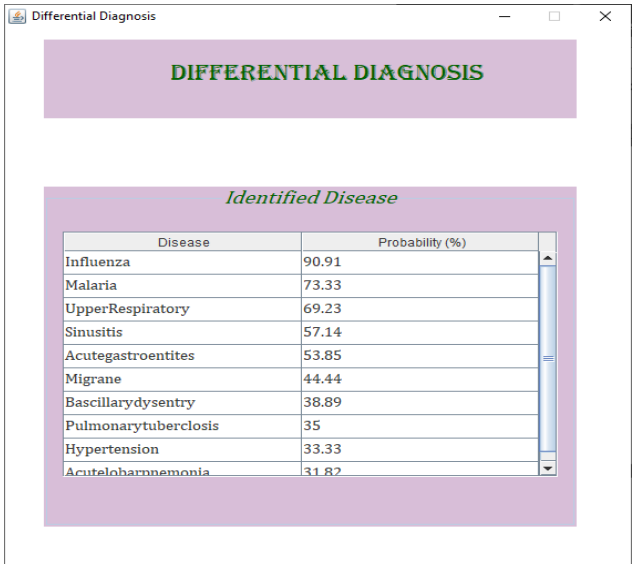

Figure 5. Output images

VI. CONCLUSION

In summary, the suggested illness prediction system is a major development in medical informatics, with its integrated modules concentrating on knowledge search, symptom analysis, pattern matching, and differential diagnosis. Through the use of machine learning techniques and historical data, the system provides healthcare providers with a potent tool for early sickness diagnosis and detection. Through the facilitation of timely intervention and individualized treatment regimens, the systematic approach not only improves the efficiency of illness management but also shows promise for improving patient outcomes. All things considered, this system has the power to completely transform healthcare procedures and bring in a new era of accurate and proactive medical treatment.

VII. FUTURE WORK

creating machine learning models that are more dependable and precise. Larger, better-quality datasets will be necessary for this, as will the creation of fresh machine learning algorithms with a focus on healthcare applications. enhancing the interpretability of machine learning models. Clinicians will be better able to trust the outcomes and comprehend how the algorithms generate their predictions as a result. creating tailored medicine-relevant machine learning models. This will entail customizing the models for each patient according to their distinct attributes, including their lifestyle, genetic makeup, and medical background.

REFERENCES

[1] Predictive Analysis for the Arbovirus Dengue using SVM Classifications, Fathima, A.S. and Manimeglai, D. (2020). International Journal of Technology and Engineering, 2: 521–527.
[2] Two feature selection methods for the Malaysian dengue epidemic detection model, Tarmizi, N.D.A. et al., 2021. Future Generation Information Technology Journals (JNIT), 4, 96–107.
[3] Analysis of Liver Disorders Using Data Mining Algorithms, Rajeswari, P. and Reena, G.S. (2019). International Journal of Technology and Computer Science, 10, 48–52.
[4] Liver Patients Classification Using Intelligent Technique, Gulia, A. et al., 2020. Journal of Computer Science and Information Technology (IJCSIT), 5, 5110–5115.
[5] Vijayarani, S. and Dhayanand, S. (2021) SVM and Naive Bayes Algorithm for Liver Disease Prediction. Science, Engineering, and Technology Researches International Journal, 4, 816–820.
[6] Intelligent Naive Bayes Approaches to Diagnose Diabetes Type-2, Sarwar, A., and Sharma, V. (2020). Issues and Challenges in Networking, Intelligences and Computing Technologies: Special Issues of International Journal of Computer Application (0975-8887), 3, 14-16.
[7] In 2019, Iyer, A., Jeyalatha, S., and Sumbaly, R. used classification mining to diagnose diabetes. Journal of Knowledge Management Process & Data Mining International (IJDKP), 5, 1–14.
[8] Tan and colleagues, A Hybrid Evolutionary Algorithm for Data Mining Attribute Selections (2021). Expert Systems with Application Journal
[9] Heart Disease Detection Using Naive Bayes Algorithms, Vembandasamy et al., 2020. Journal of Innovative Science, Engineering, and Technology - IJISET, 2, 441-444.
[10] Effective Diagnosis and Monitoring of Heart Diseases, Otoom et al., 2020. Journal of International Software Engineering and Its Applications, Volume 9, pages 143–156.