Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Enhancing Disease Forecasting Through Naïve Bayes Classification Techniques

Dr. Sharad Mathur
Computer Science
Lachoo Memorial College Sc & Tech
Jodhpur, India

Dr. Ashish Rai
Computer Science
Lachoo Memorial College Sc & Tech
Jodhpur, India

***Abstract* - A disease prediction system plays a vital role in today's healthcare landscape. Accurately predicting a disease remains a complex task due to various influencing factors. By leveraging machine learning, data analytics, and artificial intelligence, these systems assess the likelihood of medical conditions, allowing for timely and informed healthcare decisions. Such predictive approaches support early diagnosis, preventive care, tailored treatment strategies, and can significantly reduce overall medical expenses. The Naïve Bayes classifier is a probabilistic model based on Bayes' theorem, assuming that features are conditionally independent given the class label. Despite this strong assumption, it has demonstrated remarkable performance in medical applications, particularly for disease prediction. The effectiveness of the Naïve Bayes classifier lies in its ability to process medical datasets efficiently while maintaining reasonable classification accuracy. The model is particularly useful when working with small or incomplete datasets, as it requires fewer training samples compared to complex models like deep learning. Additionally, its capability to handle categorical and continuous data makes it a versatile choice for various disease prediction tasks. Among various classification algorithms, the Naïve Bayes classifier is widely used due to its simplicity, efficiency, and ability to handle uncertainty. Naïve Bayes is a widely used machine learning method that proves highly effective in predicting health outcomes, diagnosing medical conditions, and assessing future risks. It has been successfully applied in identifying the likelihood of various diseases such as diabetes, heart disease, kidney disorders, cancer, sleep apnea, and several others, making it a valuable tool in modern medical decision-making. This paper presents a comprehensive analysis of disease prediction systems based on the Naïve Bayes classifier. We discuss its theoretical foundations, advantages, limitations, and applications in diagnosing diseases such as heart disease, diabetes, and cancer..**

Keywords - Medical Dataset, Disease Diagnosis, Data Analysis, Naïve Bayes.

I. INTRODUCTION

With the increasing availability of healthcare data, data mining techniques have been widely adopted to improve disease diagnosis and prediction. Conventional diagnostic approaches often depend heavily on human judgment, which can be both time-consuming and susceptible to mistakes. In contrast, data mining plays a transformative role in predicting diseases by uncovering valuable insights from large volumes of healthcare data. It involves techniques such as classification, clustering, association rule mining, and sequential pattern analysis to detect hidden trends and relationships [1]. By identifying meaningful patterns in historical patient records, data mining helps reduce the need for unnecessary tests, speeds up diagnosis, improves accuracy, and ultimately supports better and earlier medical intervention.

Machine learning models, particularly probabilistic classifiers such as Naïve Bayes, provide a systematic approach for analyzing complex datasets and making accurate predictions. The Naïve Bayes classifier, based on Bayes' theorem, assumes conditional independence among predictor variables. Despite this strong assumption, it has demonstrated remarkable performance in medical diagnosis, making it a preferred choice for disease prediction.

This paper emphasizes the use of the Naïve Bayes classifier in predicting cancer. In recent years, India has seen a significant increase in cancer cases, earning it the informal label of the "cancer capital of the world." Among women, breast, cervical, and ovarian cancers are most commonly reported, while men are frequently diagnosed with prostate, lung, and oral cancers. The application of predictive models like Naïve Bayes offers valuable support in early detection and diagnosis of these diseases. Dona Sara Jacob et al. [2] applied classification technique and the clustering method. The outcome indicates that the classification algorithms are superior predictors than the clustering algorithms. Studies filtered all algorithms based on the Support Vector Machine lowest computing time and accuracy and it came up with the conclusion that Naïve Bayes's a superior algorithm compared SVM model. Subasini et al. examined different data mining methods used in the diagnosis and prognosis of breast cancer. Their study particularly focused on how association rule mining can be effectively applied to identify patterns that help predict the presence of breast cancer. Also it analyzes the performance of conventional supervised learning algorithms viz. C5.0, ID3, APRIORI, C4.5 and Naïve Bayes [3].

Naïve Bayes have found significant use in predicting heart disease, a condition that remains one of the top causes of death worldwide. Predictive models often take into account key factors like age, cholesterol levels, and blood pressure to assess risk. In a study by Swathi Priyadarshini et. al., clustering methods were combined with classification algorithms such as Naïve Bayes and Decision Trees to build a heart stroke prediction model. Their work highlights the potential of integrating multiple techniques to improve the accuracy and reliability of health prediction systems [4].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

II. DATA MINING FOR DISEASE PREDICTION

Data mining refers to the process of uncovering useful insights and patterns from large datasets. Common techniques include classification, clustering, prediction, association rule mining, and sequential pattern analysis. These methods are widely applied across various fields, including healthcare. In the medical domain, data mining plays a key role in disease prediction by helping to identify potential health issues without the need for excessive diagnostic tests. By reducing the number of tests required, it not only saves time but also improves the overall efficiency of the diagnostic process, ultimately leading to faster and more accurate healthcare delivery [5].

Figure 1 illustrates the step-by-step process of a disease prediction system built using a Naïve Bayes machine learning algorithm. The process begins with gathering historical data related to a specific disease, which is essential for training the model. Once the data is collected, it undergoes a cleaning phase to remove any inconsistencies or errors. After that, important features relevant to the Decision Tree model are selected. The system then proceeds through training and testing stages. Finally, based on the input patient data, the trained Naïve Bayes model generates predictions about the likelihood of the disease.

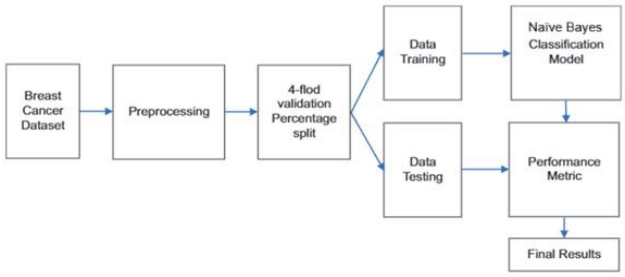

Fig. 1. Block diagram of Disease Prediction System based on Naive Bayes classifier

III. NAIVE BAYES

The Naive Bayes algorithm is a classification technique that leverages Bayesian principles, specifically Bayes' theorem, for predictive modeling [6]. Compared to other algorithms, it is less computationally demanding, making it suitable for rapidly generating mining models to uncover connections between input features and the target variable [7]. This algorithm is widely employed in the development of classifiers, which predict categorical class labels based on input data. These classifiers help determine the category to which a given input belongs.

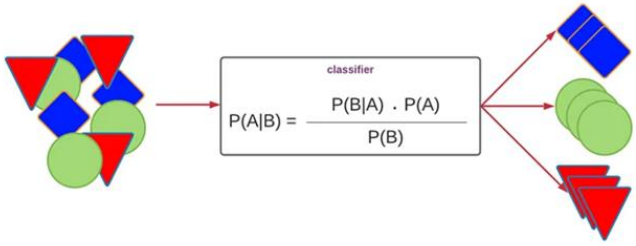

Fig.2 . Naïve Bayes classifier

To enhance the prediction framework, the proposed intelligent healthcare system utilizes a data mining technique called the "Naive Bayes classifier" [8]. This system incorporates a large dataset with numerous attributes, gathered from expert data, to accurately predict symptoms. Several artificial intelligence and data mining techniques are based on the principles of the "Naive Bayes" or "Bayes' theorem." This approach is used to develop models with predictive capabilities, leveraging available evidence to establish relationships between the target variable (dependent variable) and other influencing factors.

The Bayes theorem and Bayes approximation are both applied for Naive Bayes Method. Bayes theorem $P(c|x) = P(x|c) P(c)/P(x)$, where, $P(c|x)$ is the posterior probability of class (target) given predictor. $P(c)$ is the prior probability of class. $P(x|c)$ is the probability of predictor given class likelihood. $P(x)$ is the prior probability of predictor.

Naive Bayes is a straightforward yet powerful approach used to build models that classify data into predefined categories. Instead of being just one algorithm, it represents a group of algorithms that all rely on a shared core idea. This key assumption is that each feature in the dataset is considered independent of the others when making predictions [9]. While there are many probabilistic models available, Naive Bayes stands out for its efficiency and effectiveness in supervised learning, especially for classification problems..

IV. CANCER DISEASE PREDICTION

A recent study has raised concerns about the declining state of health in India. An international healthcare report has identified India as facing a growing crisis, referring to it as the “cancer capital of the world” due to the rapidly increasing cases of cancer and other non-communicable diseases. Projections indicate that cancer cases in India will rise from 1.46 million in 2022 to around 1.57 million by 2025—an increase expected to surpass global averages [10]. Among women, breast, cervical, and ovarian cancers are the most commonly diagnosed, while men are more often affected by prostate, oral, and lung cancers. Interestingly, research published in The Lancet Oncology highlights a reversal of global trends in India, where more women are being diagnosed with cancer than men, defying the usual 25% higher incidence seen in men worldwide.

In related research, A. S. Nath [11] explored lung cancer prediction using about 25 relevant clinical attributes. The study employed several classification algorithms, including Naive Bayes, Bayesian Networks, and J48. Among these, the Naive Bayes algorithm demonstrated the fastest model-building performance. The study further recommended enhancing the prediction system by incorporating other data mining techniques such as time series analysis, clustering, and association rule mining to improve accuracy and scope.

Rashmi et al. emphasized that breast cancer has become a significant health concern for women in developing countries, where mortality rates are steadily rising. To address this, they employed data mining techniques to analyze and evaluate classification and prediction models. Specifically, they used the Naive Bayes classification and prediction algorithms to determine whether a tumor is benign or malignant. Their study was based on the well-known Wisconsin University breast

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

cancer dataset, which was used to assess the accuracy and error rates of the applied models [12].

According to Bellaachia et al., presents an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques [13]. In the research paper, authors used SEER Public Use Data and the preprocessed data set consists of 151,886 records, available with 16 fields from the SEER database. They have analyzed the SEER data set using three data mining techniques namely Naïve Bayes, back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. Their result shows that C4.5 technique has better performance than the other two methods.

According to Kareem et al, Breast cancer, being a prevalent and common disease, highlights the significance of early detection. Perfect decision-making about breast cancer is vital for early cure and achieving positive outcomes. The percentage split assessment approach was applied, comparing performance metrics such as precision, recall, and f1-score. Kernel Naïve Bayes succeeded 100% precision in the percentage split method for breast cancer, while the Coarse Gaussian support vector machines achieved 97.2% precision in classifying breast cancer using 4-fold cross-validation [14].

V. DIABETES DISEASE PREDICTION

Insulin is a vital hormone that plays a key role in how our body processes food into energy. It helps regulate the absorption of sugars and carbohydrates, ensuring that glucose from the food we eat can enter our cells and be used as fuel. Without enough insulin—or if the body can't use it properly—glucose builds up in the bloodstream and is eventually passed out of the body through urine [15]. This condition is known as diabetes. While the exact cause of diabetes remains unclear, factors like being overweight and leading a sedentary lifestyle are known to significantly increase the risk of developing the disease.

It is estimated that around 77 million adults in India are currently living with type 2 diabetes, and an additional 25 million are in the prediabetes stage, meaning they are at high risk of developing the condition in the near future [16]. Alarmingly, more than half of those affected are unaware they have the disease, which can lead to serious health complications if not diagnosed and managed early. People with diabetes face a significantly higher risk, up to two to three times greater of experiencing heart attacks and strokes. Additionally, diabetes can lead to nerve damage (neuropathy), particularly in the feet. When this is combined with poor circulation, it can result in infections, foot ulcers, and in severe cases, may require limb amputation. Another serious complication is diabetic retinopathy, a leading cause of vision loss, caused by long-term damage to the small blood vessels in the retina due to prolonged high blood sugar levels.

To identify chances of diabetes and pre-diabetes in the American people, Wei Yu et al. evaluated two classification schemes [17].

Iyer et al. research focuses on pregnant women suffering from diabetes. In this work Naive Bayes and Support vector machine data mining techniques are applied on the PIDD dataset to forecast diabetes in a patient. Experimental performance of above three techniques are evaluated on various measures and achieved good accuracy [18].

Nongyao et al.in [19] applied algorithms which categorizes the threat of diabetes mellitus. To fulfill the objective authors has applied following well-known data mining classification techniques are Naive Bayes, Artificial Neural Networks, Logistic Regression and Decision Tree. To increase the efficiency of designed model Bagging and Boosting techniques are applied by researchers. Experimentation output depicts that Random Forest algorithm generates optimum results in between all the other techniques used.

The authors Sisodia et al. say that Naive Bayes method outperforms comparatively other algorithms. So they considered Naive Bayes classifier method as the best supervised data mining algorithm of their experiment because it generates higher accuracy in comparison to other classification algorithms with an accuracy of 76.30 % [20].

VI. HEART DISEASE PREDICTION

Heart disease has emerged as one of the leading causes of death across the globe. It can manifest through a range of symptoms such as chest pain (angina), fatigue, headaches, and swelling in the legs [21]. In many cases, the root causes are linked to lifestyle choices, including poor diet, lack of physical activity, and underlying conditions like hypertension. A major challenge in addressing heart disease is the global shortage of trained healthcare professionals, which puts additional pressure on already stretched medical systems. In this context, machine learning has shown great promise, especially in the field of disease prediction [22]. By leveraging large datasets and advanced algorithms, machine learning models can accurately assess an individual's risk of developing heart disease, supporting early diagnosis and timely intervention.

Using a naïve Bayes classifier, a mining technique model was designed to detect the knowledge pertaining to the cardiovascular disease profile and the degree of cardiovascular disease risk factors for adults based on the medical record. Two techniques were used to evaluate this study: a calculation of accuracy, sensitivity, and specificity as well as a consultation with internists and cardiologists. These include kidney function, coronary artery function, blood cholesterol levels, and diabetes mellitus. The values of these factors—risk level 1, risk level 2, and risk level 3—were used to establish class labels [23].

Gudadhe et al. demonstrated that Naïve Bayes classifiers could achieve high accuracy rates when combined with feature selection techniques. By reducing irrelevant features, the model's performance significantly improved [24]. Anbarasi et al. applied a hybrid model combining Naïve Bayes with genetic algorithms for feature selection, resulting in enhanced predictive accuracy and reduced computational overhead [25]. Rajkumar & Reena explored the use of Naïve Bayes alongside other classifiers for heart disease prediction, finding that Naïve Bayes performed well, particularly in datasets with smaller sample sizes [26]. Kumar & Samanta (2012) emphasized the interpretability of Naïve Bayes models, making them particularly suitable for medical applications where understanding the reasoning behind predictions is critical [27]. Khan et al. proposed an ensemble approach combining Naïve

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Bayes with other classifiers, which enhanced overall predictive performance for heart disease detection [28]. Dey et al. evaluated Naïve Bayes in combination with data preprocessing techniques such as normalization and discretization, which improved the model's accuracy [29]. Patel et al. conducted a comparative study on various classifiers, showing that while Naïve Bayes was not the top performer in accuracy, its simplicity and speed made it a viable option for initial screening processes [30]. Haider et al. conducted a comprehensive analysis of various machine learning algorithms for cardiovascular disease prediction, reaffirming the role of Naïve Bayes as a strong baseline model [31].

VII. CONCLUSION

Disease prediction systems have become an essential component of modern healthcare, as they enable the early identification of individuals who are at elevated risk for developing specific medical conditions. Early risk detection supports healthcare professionals in designing personalized treatment plans, anticipating future clinical needs, and implementing long-term management strategies that contribute to improved patient outcomes. Moreover, these systems play a significant role in reducing hospital readmission rates by facilitating timely and targeted interventions for patients who may be prone to post-discharge complications. The Naïve Bayes classifier plays a crucial role in disease prediction due to its simplicity and efficiency. While it has some limitations, ongoing research has led to significant improvements in its predictive capabilities. To overcome the limitations of the Naïve Bayes classifier, researchers have explored hybrid models that combine it with decision trees, neural networks, and ensemble learning techniques. Feature selection methods have also been developed to improve classification accuracy by eliminating irrelevant attributes. The classifier remains a valuable tool for disease diagnosis and clinical decision support, making it an essential component of modern healthcare informatics. Future advancements in artificial intelligence and big data analytics are expected to further refine its applications in medical science. The integration of deep learning and adaptive learning approaches could further enhance the model's performance in healthcare applications.

REFERENCES

[1] K. Aftarczuk, "Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems," Blekinge, 2007.

[2] Dona Sara Jacob, Rakhi Viswan, V Manju, L PadmaSuresh,Shine Raj, "A Survey on Breast Cancer Prediction Using Data Mining Techniques", IEEE Access, 2018, ISBN: 978-1-5386-3479-0.

[3] Mathur S., Rai A., Mathur D., "Predictive Data Mining for Disease Diagnosis-Decision Tree Approach", "INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)", ISSN:2278-0181, Vol 14,no. 05, May 2025.

[4] B. Sethuraman and S. Niveditha, "Cerebrovascular Accident Prognosis using Supervised Machine Learning Algorithms," in 2023 World Conference on Communication & Computing WCONF, RAIPUR, India: IEEE, pp. 1–8. doi:10.1109/WCONF58270.2023.10235122.

[5] Venkatesh, K., et al. "Identification of Disease Prediction Based on Symptoms Using Machine Learning." JAC: A Journal Of Composition Theory 14.6 (2021).

[6] Mathur S., Joshi B., "Application of Naïve Bayes Classification for disease Prediction", International Journal of Management, IT & Engineering, Vol 9(4), 80-87, April 2019, ISSN: 2249-0558.

[7] Amit Kumar Das, Aman Kedia,"Data mining techniques in Indian Healthcare: A Short Review", 2015 International Conference on Man and Machine Interfacing (MAMI), 978-1-5090-0225-2/15.

[8] Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. In: Liu, C., Wang, L., Yang, A. (eds) Information Computing and Applications. ICICA 2012. Communications in Computer and Information Science, vol 308. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34041-3_27

[9] W. Song, C. H. Li and S. C. Park, "Expert Systems with Applications Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures", Expert Syst. Appl, vol. 36, no. 5, pp. 9090-9110, 2009.

[10] Sathishkumar, Krishnan; Chaturvedi, Meesha; Das, Priyanka; Stephen, S.; Mathur, Prashant. Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India. Indian Journal of Medical Research 156(4&5):p 598-607, Oct–Nov 2022. | DOI: 10.4103/ijmr.ijmr_1821_22

[11] A.S. Nath, A. Pal, S. Mukhopadhyay, and K.C. Mondal, "A survey on cancer prediction and detection with data analysis", Innov. Syst. Softw. Eng., vol. 16, no. 3, pp. 231-243, 2019.

[12] Ms. Rashmi G D, Mrs. A Lekha, Dr. Neelam Bawane, "Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset", IEEE Access, 2015, pg. no.108-109

[13] Mathur S., Joshi B., Survey on Needs, "Applications and Agorithms of Data Mining for Healthcare", International Journal of Advance Research in Science and Engineering, Vol 6(8), 80-87, August, 2017, ISSN: 2319-8346.

[14] Thikra Ali Kareem, Muzhir Shaban Al-Ani, Salwa Mohammed Nejres, "Efficient Breast Cancer Dataset Analysis Based on Adaptive Classifiers", UHD Journal of Science and Technology, Jan 2024 ,Vol 8, Issue 1.

[15] T. Santhanam and M. S. Padmavathi, "Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," Procedia Comput. Sci., vol. 47, no. C, pp. 76–83, 2014.

[16] Mathur S., Joshi B.,"Study on Data Mining Techniques and Applications in Healthcare", Advances in Computer Science and Information Technology (ACSIT), 2017, 4(4),ISSN: 2393-9907.

[17] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Medical Informatics Decision Making 10, Artical number 16, (2010).

[18] Iyer, A., S, J.,Sumbaly,R.,2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14.doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.

[19] Nongyao Nai-arun,N.,Moungmai,R.,2015.Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014.

[20] Sisodia, Sisodia, 2018, Prediction of Diabetes using Classification Algorithms, International Conference on Computational Intelligence and Data Science (ICCIDS2018), Procedia Computer Science 132 (2018) 1578–1585.

[21] Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Ramesh Thelkar, A. (2022). Implementation of a heart disease risk prediction model using machine learning. Computational and Mathematical Methods in Medicine,2022,1–14. https://doi.org/10.1155/2022/6517716.

[22] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. International Journalof Engineering Research & Technology,9(4), 659–662. http://doi.org/10.17577/IJERTV9IS040614

[23] Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Healthc Inform Res. 2016 Jul;22(3):196-205. doi: 10.4258/hir.2016.22.3.196. Epub 2016 Jul 31. PMID: 27525161; PMCID: PMC4981580.

[24] Gudadhe, M., Wankhade, K., & Dongre, S. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. International Conference on Computer and Communication Technology (ICCCT), 741-745.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

[25] Anbarasi, M., Anupriya, E., & Iyengar, N.C.S.N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376.

[26] Rajkumar, A., & Reena, G.S. (2010). Diagnosis of heart disease using data mining algorithm. Global Journal of Computer Science and Technology, 10(10), 38-43.

[27] Kumar, A., & Samanta, S. (2012). A machine learning approach for the diagnosis of heart disease. International Journal of Engineering Research and Applications, 2(4), 2035-2039.

[28] Khan, M.A., Algarni, A.D., & Nayak, R.S. (2019). Machine learning-based heart disease prediction using Naïve Bayes and ensemble methods. Procedia Computer Science, 167, 906-915.

[29] Dey, S., Bhattacharya, S., & Nath, A. (2016). Heart disease prediction using data mining techniques. International Journal of Modern Engineering Research, 6(2), 68-73.

[30] [30] Patel, J., Tejal, K., & Sharma, S. (2015). A comparative study of predictive data mining techniques. International Journal of Computer Applications, 113(7), 1-5.

[31] Haider, S., et al. (2020). A comprehensive analysis of machine learning algorithms for cardiovascular disease prediction. Journal of Healthcare Engineering.