A comparitive analysis for Genetic Disease Detection Accuracy Through Machine Learning Models on Datasets

Nimay Seth[1], Harshita Kaushik[2] , Khushboo Sharma[3] , Hemant Sharma[4]

[1]IBDP Student, Jayshree Periwal International School, Jaipur, Rajasthan, India
[2]Assistant Professor, Department of Computer Science & Engineering VGU, Jaipur, Rajasthan, India
[3]Assistant Professor, Department of Electrical Engineering VGU, Jaipur, Rajasthan, India
[4]Reaearch Scholar, Department of Computer Science and Engineering, IIIT Kota, Kota, Rajasthan India

ABSTRACT

Over the past few decades, increasingly hectic lifestyles and careless behaviors have contributed to a rise in life-threatening conditions, with genetic diseases emerging as one of the leading causes of death worldwide. The ability to accurately and quickly predict genetic diseases is vital for effective prevention. While various methods have been developed to support healthcare professionals, each algorithm comes with its own unique challenges. This paper presents a comparative study focused on improving prediction accuracy while reducing false alarms. The study evaluates the performance of individual algorithms, including Hidden Markov Model (HMM), Naive Bayes (NB), and Multilayer Perceptron (MLP). The results demonstrate that, when compared to each other, these algorithms show varying levels of effectiveness, with certain approaches offering notable enhancements in prediction accuracy.

Keywords—Data mining, Naive bayes (NB), Neural Network (NN), Hidden Markov Model (HMM), Multilayer Perceptron (MLP)

INTRODUCTION

Genetic diseases have been the leading cause of death globally over the past decade. To assist healthcare professionals in diagnosing genetic diseases, researchers have employed various data mining techniques. Nowadays, the healthcare industry generates vast amounts of data related to patient diagnoses, forming extensive biomedical datasets. Many healthcare organizations utilize intelligent healthcare information systems to extract and analyze these datasets, uncovering hidden patterns or relationships within the data. Data mining offers a range of techniques to discover valuable knowledge from medical datasets.

The primary goal of healthcare organizations is to deliver high-quality services at an affordable cost. High-quality service entails accurately diagnosing diseases and providing effective treatments for patients. Analyzing genetic disease patient databases can be likened to real-world applications, where a doctor’s expertise plays a crucial role in assigning appropriate weights to each attribute based on their impact on disease prediction. Attributes with a higher impact are given more weight, providing healthcare professionals with additional knowledge to support decision-making.

Numerous factors contribute to genetic diseases, including poor dietary habits, stress, lack of exercise, high blood pressure, smoking, alcohol consumption, drug abuse, cholesterol levels, and elevated blood sugar. Consuming fatty foods can weaken blood vessels, leading to various conditions. Increased pressure on arteries can thicken the heart walls, obstructing blood flow and potentially causing complications.

To simplify the complexity of diagnosing genetic diseases, this paper introduces a novel method by applying a new ensemble classification technique in data mining.

HUMAN BODY STRUCTURE & CAUSING FACTORS OF GENETIC DISEASES

- **Age:** Over 83% of individuals who die from coronary heart disease are 65 or older. Similarly, the risk of diabetes increases with age. Older women are more likely to die of heart attacks within a few weeks of the event compared to older men.

- **Gender:** Men are at a higher risk of heart attacks than women, often experiencing them earlier in life. However, after menopause, the risk of heart disease in women increases, though it remains lower than in men. Gender differences also exist in the prevalence and management of diabetes.
- **Family History:** A family history of heart disease or diabetes significantly raises the likelihood of developing these conditions. Individuals with parents or close relatives who have these diseases are at a higher risk.
- **Ethnicity:** Certain ethnic groups, such as African Americans, Mexican Americans, American Indians, Native Hawaiians, and some Asian Americans, face a higher risk of heart disease and diabetes compared to Caucasians.
- **Smoking:** Smoking increases the risk of developing heart and diabetes disease by two to four times. It also contributes to the development of diabetes, further elevating the risk of cardiovascular complications.
- **High Cholesterol:** Elevated blood cholesterol levels increase the risk of heart and diabetes disease, which is also closely linked with diabetes.
- • **Elevated Blood Pressure:** Elevated blood pressure intensifies the workload on the heart, leading to thickening and hardening of the heart muscle. This condition heightens the likelihood of stroke, heart attack, kidney failure, and congestive heart failure. The risk of a heart attack or stroke increases substantially when high blood pressure is paired with smoking, high cholesterol, or diabetes.
- • **Lack of Physical Activity:** Lack of physical activity is a contributing factor to both coronary heart disease and diabetes.
- • **Impact of Diabetes on Heart Health:** The presence of diabetes significantly boosts the risk of developing cardiovascular diseases. Around 75% of individuals with diabetes succumb to some type of heart or vascular disease.
- **Excess Weight:** Individuals with excess body fat, particularly around the waist, are more prone to developing heart disease and stroke, even in the absence of other risk factors. Obesity is also a major contributor to the development of diabetes.

These risk factors are critical when analyzing heart and diabetes datasets, as they provide insights into the prevalence and management of these conditions across different populations.

RELATED WORK

An Intelligent Prediction System for Heart Disease and Diabetes (IPSHDD) was developed using data mining techniques such as Hidden Markov Model (HMM), Multilayer Perceptron (MLP), and Naive Bayes (NB). The study highlighted the unique strengths of each methodology in meeting specific data mining objectives. IPSHDD demonstrated its ability to address complex queries that traditional decision support systems could not handle, enabling the discovery of critical insights, such as patterns and relationships among medical factors related to heart disease and diabetes. The system is designed to be web-based, user-friendly, scalable, reliable, and expandable, making it a robust tool for healthcare professionals.

In another study, the authors (Sharma et al., 2021) [4] introduced a novel prediction mechanism utilizing both linear and nonlinear features of Heart Rate Variability (HRV) for heart and diabetes datasets. They applied statistical and classification techniques to develop a multi-parametric feature of HRV and evaluated its linear and nonlinear properties in three different positions: supine, left lateral, and right lateral. The study assessed various classifiers, including Naive Bayes (NB), Hidden Markov Model (HMM), and Multilayer Perceptron (MLP), with MLP often outperforming the other methods.

Srinivas et al., 2010 [2] proposed a method for predicting heart disease, blood pressure, and sugar levels using neural networks. They conducted experiments on patient records, training and testing the Neural Network with 13 input variables such as age, blood pressure, and angiography reports. The supervised network, trained with the back propagation algorithm, was recommended for diagnosing heart diseases and diabetes. When new data was entered by a healthcare provider, the system compared it to the trained data to generate a list of potential diseases the patient might be at risk for.

Shouva et al., 2021 [9] developed a decision tree using patient data to predict survival outcomes after out-of-hospital cardiac arrest. Their study demonstrated the significant contribution of data mining methods like HMM and NB in sorting variables and understanding the data's impact on the study's outcomes. However, a significant limitation was the extensive data collection required to create an accurate model.

Sharma et al., 2021 [5] proposed an associative classification approach for diagnosing cardiovascular disease by assessing HRV from ECG data, pre-processing it, and identifying heart disease patterns. They conducted experiments using a dataset of 670 people, divided into two groups: normal individuals and those with heart disease. The comparative study highlighted the strengths and weaknesses of HMM, MLP, and NB when applied to this dataset.

Princy et al., 2020 [10] explored the problem of identifying constrained association rules for predicting heart disease within the context of genetic and metabolic factors. The study introduced three constraints to reduce the number of patterns and running time of the experiments, which were used to predict the presence or absence of heart disease and diabetes in specific populations.

Princy et al., 2020 [10] proposed a novel heuristic for the efficient computation of sparse kernels in SUPANOVA, which was applied to datasets related to heart disease and diabetes, as well as a benchmark Boston housing market dataset. This method, leveraging MLP and NB algorithms, achieved an 83.7% prediction accuracy, outperforming results obtained through Support Vector Machines and equivalent kernels.

Rahmani et al., 2021 [8] introduced a coactive neuro-fuzzy inference system (CANFIS) for predicting heart disease and diabetes. The CANFIS model combined the adaptive capabilities of neural networks with fuzzy logic, integrated with a genetic algorithm. The model's performance, evaluated based on training results and classification accuracies, showed promise in predicting both heart disease and diabetes when compared to HMM and NB.

Jabbar et al., 2018 [17] conducted an empirical study on predicting heart disease using classification data mining techniques, specifically comparing the performance of Decision Trees (DT), Naive Bayes (NB), Hidden Markov Model (HMM), and Multilayer Perceptron (MLP) on heart disease and diabetes datasets. The large volume of data required dimensionality reduction using attribute selection methods before classification. The study found that the NB classifier achieved better accuracy for heart disease prediction after applying the CFS attribute selection method, while MLP showed strong performance in diabetes prediction.

Princy et al., 2020[10] developed a predictive model for coronary heart disease using a decision tree algorithm, which they compared against HMM and NB within a case-control study. The authors built a decision tree using traditional CHD-related factors and evaluated it on a testing group of 706 records, achieving a 94% accuracy rate, surpassing the results of HMM and NB in the study.

Princy et al., 2020 [10] proposed an efficient classification approach to categorize ECG signals into different arrhythmias to aid in diagnosing heart disease and related diabetes complications. The ECG signals were preprocessed using a Morphological Filter, and the extracted features were classified using the PBNN network. The system was evaluated with the MIT-BIH arrhythmia database, comparing its performance against HMM and NB.

Sharma et al., 2021 [4] developed a hybrid classification system for diagnosing heart disease and diabetes using the ReliefF and Rough Set (RFRS) methods. This system was evaluated in a comparative study against HMM, MLP, and NB, showcasing the relative strengths and weaknesses of each approach in these complex prediction tasks.

PROPOSED SCHEME

To enhance the accuracy of genetic disease prediction, this paper conducts a comparative study of individual classification algorithms. The study evaluates the predictive power of Naïve Bayes (NB), Multilayer Perceptron (MLP), and Hidden Markov Model (HMM), three widely recognized data mining techniques. Classification algorithms typically involve selecting the most suitable hypothesis from a set of alternatives that align with a set of observations. The data classification process consists of two main steps: the first involves constructing a classifier that defines a predetermined set of data classes or concepts, known as the learning or training phase. During this phase, each classification algorithm builds its classifier by analyzing a training set composed of database tuples and their associated class labels. The process flow diagram of this comparative study is illustrated in Fig 1.

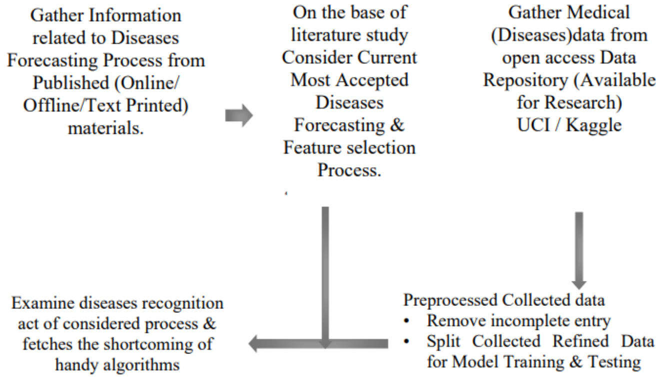
Gather Information related to Diseases Forecasting Process from Published (Online/ Offline/Text Printed) materials.
On the base of literature study Consider Current Most Accepted Diseases Forecasting & Feature selection Process.
Gather Medical (Diseases)data from open access Data Repository (Available for Research) UCI / Kaggle
Examine diseases recognition act of considered process & fetches the shortcoming of handy algorithms
Preprocessed Collected data
• Remove incomplete entry
• Split Collected Refined Data for Model Training & Testing

Fig. 1 Process Flow Diagram of Proposed Work

Based on the figure provided, the proposed method begins by gathering information related to disease forecasting processes, specifically for heart disease and diabetes, from various sources, including published materials available online, offline, or in text format. The literature study then focuses on identifying the most accepted forecasting and feature selection processes for these diseases. Next, medical data related to heart disease and diabetes is collected from open-access repositories such as UCI and Kaggle.

The collected data undergoes preprocessing, where incomplete entries are removed, and the refined data is split for model training and testing. The subsequent steps involve examining the effectiveness of disease recognition through the considered processes and addressing any shortcomings in the individual algorithms used.

In this comparative study, Naive Bayes (NB) is first applied as the initial classification technique on the selected features. If NB accurately predicts the disease, the correctly predicted instances are removed from the dataset. However, if NB fails to make an accurate prediction, the dataset is passed to the next layer, where the Hidden Markov Model (HMM) or Multilayer Perceptron (MLP) is used for further prediction.

EXPERIMENTAL SETUP & RESULT ANALYSIS

To evaluate the performance of the designed approach against both traditional methods and other recently proposed techniques, a series of diverse simulations were conducted using two different datasets: Heart dataset and Diabetes dataset.

Table 1 Dataset Used For Experiments

S.No.	Dataset	No. of Attributes	Instances
1.	Heart dataset	14	297
2.	Diabetes dataset	33	660

A. Performance Analysis Using the Heart Dataset

In this phase of the experiment, the dataset was utilized, which includes 14 attributes and 297 instances. Each instance typically represents the record of an individual, either with heart disease or without. The dataset contains records for 120 individuals diagnosed with heart disease and 177 records for individuals without the condition.

Table 2 Comparative Performances over Heart Dataset

S.No.	Prediction Techniques	Accuracy (%)
1	Hidden Markov Model (HMM)	85.149
2	Naïve Bayes (NB)	84.489
3	Multi-Level Perceptron (MLP)	83.499

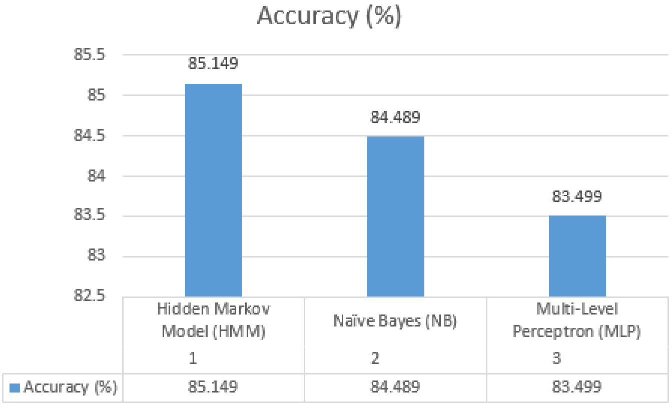

Fig. 2 Performance of Individual Algorithms (HMM, NB, MLP) Compared with Heart Dataset

The figure above illustrates the comparative performance of individual algorithms, including Hidden Markov Model (HMM), Naïve Bayes (NB), and Multilayer Perceptron (MLP), when applied to the Heart dataset. The results indicate that certain algorithms consistently deliver superior evaluation outcomes compared to others, in line with their performance in previous assessments.

B. Performance Analysis of the Proposed Scheme Using Diabetes Datasets

To further evaluate the efficiency of the designed approach under real-time conditions and varying parameters, an additional experiment was conducted using diabetes datasets. The results, presented in Table 3, offer a comparative analysis similar to the previous evaluation. The statistics reveal that, consistent with earlier findings, the proposed data prediction technique outperforms existing standalone prediction methods. The evaluation results confirm that the designed algorithm provides real-time efficiency for diabetes prediction systems.

Table 3 Performances over Diabetes Dataset

S.No.	Prediction Techniques	Accuracy (%)
1	Hidden Markov Model (HMM)	77.96
2	Naïve Bayes (NB)	82.04
3	Multi-Level Perceptron (MLP)	80.03

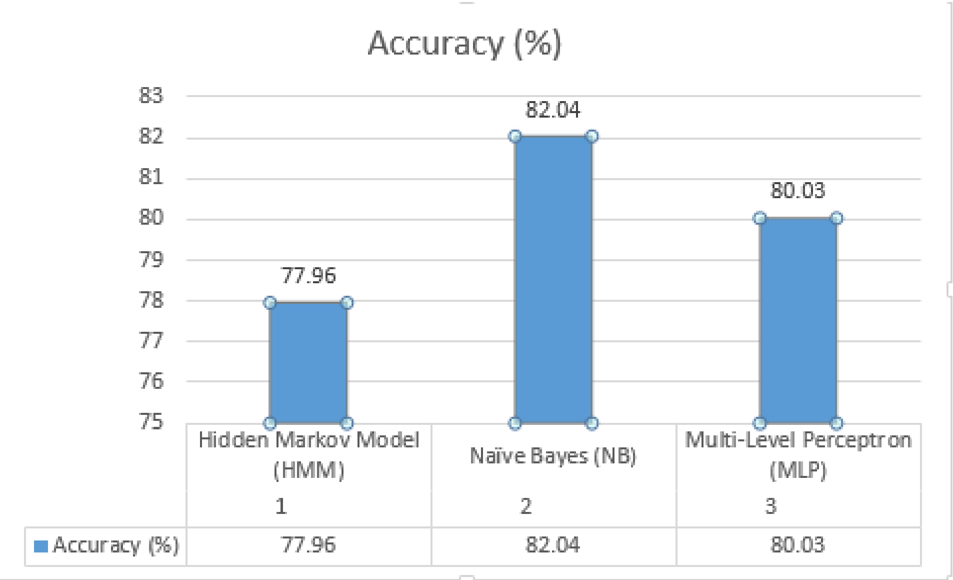

Fig. 3 Performance of Individual Algorithms Compared to Standalone Prediction Methods with Diabetes Dataset

The figure above demonstrates that, in the context of the Diabetes dataset, the comparative study of individual algorithms—Hidden Markov Model (HMM), Naive Bayes (NB), and Multilayer Perceptron (MLP)—reveals that these algorithms perform significantly better than standalone prediction methods. This finding is consistent with the performance observed in previous assessments, highlighting the advantages of these individual approaches in predicting diabetes outcomes.

CONCLUSION & FUTURE WORK SCOPE

This investigation presents a comparative study that examines the predictive capabilities of individual data mining algorithms, specifically Hidden Markov Model (HMM), Naïve Bayes (NB), and Multilayer Perceptron (MLP). The study not only compares the performance of these algorithms in data prediction but also incorporates an attribute selection process to optimize their effectiveness. Datasets were sourced from the publicly available UCI data repository to validate the efficiency of each algorithm. The evaluation results consistently demonstrate the varying levels of effectiveness and efficiency of these individual approaches in data prediction.

The graphs indicate that the Hidden Markov Model (HMM) consistently outperforms the other models, achieving 85.149% accuracy in the first experiment and 77.96% in the second. The Naive Bayes (NB) model displays significant variability, with 84.489% accuracy in the first test and an increase to 82.04% in the second. Meanwhile, the Multi-Level Perceptron (MLP) shows stable performance, recording 83.499% in the first experiment and a slight drop to 80.03% in the second, making it generally less effective than the HMM.

Future research could explore ways to reduce the time required for data evaluation, potentially enhancing the performance of these individual algorithms in new and innovative ways.

REFERENCES

[1]. Shakuntala Jatav and Vivek Sharma “An Algorithm for Predictive Data Mining Approach in Medical Diagnosis” International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018, pp.-11-20.

[2]. K.Srinivas B.Kavihta Rani Dr. A.Govrdhan " Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

[3]. P. Sharma, S. Saxena and Y. Mohan Sharma, "An Efficient Decision Support Model Based on Ensemble Framework of Data Mining Features Assortment & Classification Process," 2018 3rd International Conference on Communication and Electronics Systems(ICCES), 2018, pp. 487-491

[4]. Sharma, Y.M., Saini, P.K., Shalini, Sharma, N. (2021). Effective Decision Support Scheme Using Hybrid Supervised Machine Learning Procedure. Proceedings of the Second International Conference on Information Management and Machine Intelligence. Lecture Notes in Networks and Systems, vol 166. Springer, Singapore. https://doi.org/10.1007/978-981-15-9689-6_61

[5]. Jaiswal, O., Saini, P.K., Shalini, Sharma, Y.M. (2021). Analyze Classification Act of Data Mining Schemes. In: Goyal, D., Gupta, A.K., Piuri, V., Ganzha, M., Paprzycki,

[6]. M. (eds) Proceedings of the Second International Conference on Information Management and Machine Intelligence. Lecture Notes in Networks and Systems, vol 166. Springer, Singapore. https://doi.org/10.1007/978-981-15-9689-6_54

[7]. Amir Masoud Rahmani, Efat Yousefpoor, Mohammad Sadegh Yousefpoor, Zahid Mehmood, Amir Haider, Mehdi Hosseinzadeh and Rizwan Ali Naqvi “Machine Learning (ML) in Medicine: Review, Applications, and Challenges” Mathematics 2021, 9, 2970.

[8]. Weissler, E.H., Naumann, T., Andersson, T. et al. The role of machine learning in clinical research: transforming the future of evidence generation. Trials 22,537 (2021).

[9]. Rahmani, Amir Masoud, Efat Yousefpoor, Mohammad Sadegh Yousefpoor, Zahid Mehmood, Amir Haider, Mehdi Hosseinzadeh, and Rizwan Ali Naqvi. 2021. "Machine Learning (ML) in Medicine: Review, Applications, and Challenges" Mathematics 9, 22: 2970.

[10]. Shouva, R.; Fein, J.A.; Savani, B.; Mohty, M.; Nagler, A. Machine learning and artificial intelligence in haematology. Br. J. Haematol. 2021, 192, 239–250.

[11]. R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570-575.

[12]. C. -H. Lin, P. -K. Yang, Y. -C. Lin and P. -K. Fu, "On Machine Learning Models for Heart Disease Diagnosis," 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2020, pp. 158-161.

[13]. Alafif, T.; Tehame, A.M.; Bajaba, S.; Barnawi, A.; Zia,

[14]. S. Machine and deep learning towards covid-19 diagnosis and treatment: Survey, challenges, and future directions. Int. J. Environ. Res. Public Health 2021, 18, 1117.

[15]. Tayarani-N, M.-H. Applications of artificial intelligencein battling against covid-19: A literature review. Chaos Solitons Fractals 2020, 110338.

[16]. Smiti, A. When machine learning meets medical world: Current status and future challenges. Comput. Sci. Rev. 2020, 37, 100280.

[17]. Madhukar NS, Khade PK, Huang L, Gayvert K, Galletti G, Stogniew M, et al. A Bayesian machine learning approach for drug target identification using Weissler et al. Trials (2021) 22:537 Page 12 of 15 diverse data types.Nat Commun. 2019;10(1):5221.

[18]. Shalini, Saini, P.K., Sharma, Y.M. (2021). An Intelligent Hybrid Model for Forecasting of Heart and Diabetes Diseases with SMO and ANN. In: Shorif Uddin, M., Sharma, A., Agarwal, K.L., Saraswat, M. (eds) Intelligent Energy Management Technologies. Algorithms for Intelligent Systems. Springer, Singapore.

[19]. M. A. Jabbar, Shirina Samreen, Rajanikanth Aluvalu “The Future of Health care: Machine Learning” International Journal of Engineering & Technology, 7 (4.6) (2018) 23-25

[20]. K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 910-914.

[21]. X. Wenxin, "Heart Disease Prediction Model Based on Model Ensemble," 2020 3rd International Conference onArtificial Intelligence and Big Data (ICAIBD), 2020, pp.195-199,

[22]. Hong, W.H., Yap, J.H., Selvachandran, G. et al. Forecasting mortality rates using hybrid Lee–Carter model, artificial neural network and random forest

[23]. . Complex Intell. Syst. 7, 163–189 (2021).

[24]. S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-7

[25]. C. S. Prakash, M. Madhu Bala and A. Rudra, "Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-4,