FOR ANALYSING AND IMPROVE THE ACCURACY OF HEART DISEASE PREDICTION USING MACHINE LEARING TECHNIQUE

Abstract

The field of medical science has created significant attention from researchers due to emergency need to address the different causes of early mortality in humans. Among these causes, heart-related illnesses have emerged as a prominent concern. Numerous investigators have delved into the study of heart diseases, aiming to develop effective methods for preserving human life and aiding healthcare professionals in their recognition, prevention, and management. While several approaches have been proposed, each with its own limitations, there is a growing interest in utilizing decision tree analysis, random forest, and hybrid models to tackle this problem. This proposed approach entails a robust analysis of these methods for selecting the most appropriate to improve the predication of the disease. The analysis involves executing different feature sets with each method and examining the resulting statistics. Through this research, a smart and effective method will be developed to enhance decision-making capabilities for healthcare experts in the realm of heart disease management.

Keywords: Robust, Prevention, Decision Tree, Random Forest XG boost Stacking Classifier, Numerous, investigators

Authors

Dr. Srinivasan Jagannathan
Assistant Professor
Department of Computer Applications
Madanapalle Institute of Technology & Science
Madanapalle, Andhra Pradesh, India.
drsrinivasanj@mits.ac.in

I. INTRODUCTION

In recent years, researchers have made significant progress in developing methods to improve the quality of systems and enhance their performance. However, many of these methods fail to cater to the specific needs and characteristics of professionals working in different domains. This limitation necessitates further research and the design and implementation of methods that can better support professionals in their work. This main aim of the paper is to address this gap by presenting a novel approach that utilizes data mining methods to enhance the quality of service in systems. The paper begins by discussing the research objectives, motivation, and key findings. It highlights the importance of data mining as an innovative tool with great potential for companies to extract valuable information from their vast data repositories. The main DM concept of Knowledge Discovery in Databases, is the process of unknown and predictive information from large amounts of datasets It plays major roles in identifying valid, unknown, potentially useful, and understandable patterns in data. Ensemble learning techniques have gained significant attention from researchers in various fields in healthcare, finance, insurance, manufacturing, and bioinformatics. These techniques involve combining multiple classifiers to improve the classification of new test samples. Research has demonstrated that ensemble approaches can outperform individual models and reduce generalized error rates. The main object of using ensemble approaches to build a model that combines diverse individual classifier techniques to achieve higher accuracy. In the context of this paper to focus on the field of bioinformatics due to its popularity in research. The paper describes several ensembles learning methods, including Bagging, Boosting, and Stacking. Bagging involves creating multiple bags or samples from the original dataset to train separate classifiers, which are then combined to make predictions. Boosting, on the other hand, sequentially learns from the complete dataset and adjusts the weights of training data points based on their classification accuracy. Stacking involves combining different predictors to generate improved results through two phases of learning. While there has been extensive research in the field of ensemble learning, this chapter presents a comparative study of different ensemble techniques and their application in bioinformatics. Previous research studies have demonstrated the superiority of ensemble learning approaches over individual machine learning algorithms for tasks such as breast cancer metastasis classification and EEG signal classification. Overall, this chapter aims to contribute to the field of improving system QoS by introducing novel ensemble learning techniques and demonstrating their effectiveness in the context of bioinformatics. By leveraging data mining methods and ensemble learning, professionals can benefit from more accurate and reliable systems tailored to their specific needs.

II. RELATED WORKS

1. **Comparative study of Classification Techniques in Data Mining Using Different Datasets:** Data mining plays a major role in extracting valuable insights and patterns from huge amounts of data, the established organizations to make informed decisions. Classification, involves categorizing data instances into predefined classes or groups based on their features. Various classification techniques have been developed and employed to tackle diverse real-world problems. This paper aims to provide a comparative analysis of different classification techniques in data mining using a range of datasets. Understanding the strengths and weaknesses of various classification algorithms is essential for selecting the most suitable technique for a given dataset and problem

domain. This analysis will explore popular classification algorithms, including decision trees, support vector machines (SVM), k-nearest neighbors (KNN), random forests, and neural networks. Furthermore, to ensure the robustness of the analysis, different datasets from various domains will be utilized. These datasets may include medical records, customer demographics, financial data, and more. By employing diverse datasets, this analysis aims to evaluate the performance, accuracy, efficiency, and interpretability of each classification technique. The findings from this comparative analysis will contribute to the understanding of classification techniques' effectiveness in data mining and aid in identifying the most appropriate algorithms for specific datasets and applications. This research is valuable for practitioners and researchers seeking insights into the comparative performance of classification techniques and their applicability to real-world problems.

2. **Prediction of Occupational Accidents Using Decision Tree Approach" IEEE Annual India Conference (INDICON):** Occupational accidents pose significant challenges to the safety and well-being of workers across various industries. Effective prediction of such accidents can play a vital role in mitigating risks, preventing injuries, and improving overall workplace safety. In recent years, machine learning algorithms have emerged as powerful tools for analyzing and predicting complex patterns in various domains, including occupational safety. This paper focuses on the application of a Decision Tree approach to predict occupational accidents. One of the Decision Trees algorithms is popular machine learning algorithms that use a tree-like model of decisions and their possible consequences. They are widely used for classification and regression tasks due to their simplicity and interpretability. The main objective of this paper is to develop a predictive model that can accurately forecast the occurrence of occupational accidents based on relevant input variables. The proposed model will consider various factors such as workplace environment, employee demographics, previous accident records, and safety measures implemented by the organization. By identifying the significant risk factors, employers can proactively implement preventive measures and allocate resources effectively to minimize the occurrence of accidents. The analysis and prediction of occupational accidents using a decision Tree approach can provide valuable insights to employers, safety professionals, and policymakers for designing targeted interventions and strategies to enhance workplace safety. This research aims to contribute to the existing body of knowledge on accident prevention and create a safer working environment for employees in India and beyond.

3. **A Comparative Study of Ensemble Learning Methods for Classification in Bioinformatics:** Ensemble learning method have gained significant attention in the field of bioinformatics due to their ability to improve classification accuracy by combining multiple learning models. This comparative study aims to explore and evaluate various ensemble learning techniques for classification tasks in bioinformatics. The integrating diverse base classifiers, such as decision trees, support vector machines, and neural networks, ensembles can leverage their collective intelligence to handle complex biological data sets. Through a comprehensive analysis, this study seeks to identify the most effective ensemble methods, considering their performance metrics, computational efficiency, and robustness in dealing with the challenges posed by bioinformatics datasets. This research provides valuable insights for researchers and practitioners in the

bioinformatics community, aiding in the selection and application of ensemble learning approaches for improved classification outcomes.

4. **A Novel Paradigm of Melanoma Diagnosis Using Machine Learning and Information Theory:** Melanoma, the deadliest form of skin cancer, poses a significant public health challenge worldwide. Traditional methods of melanoma diagnosis heavily rely on visual inspection by dermatologists, which can be subjective and prone to human error. To address this issue, a novel paradigm integrating machine learning and information theory has emerged as a promising approach. By leveraging advanced algorithms and the power of data analysis, this paradigm aims to revolutionize melanoma diagnosis by extracting valuable insights from large datasets. Through the integration of machine learning techniques and information theory principles, this innovative approach holds the potential to enhance accuracy, efficiency, and early detection of melanoma, ultimately improving patient outcomes and saving lives.

5. **Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data:** In recent years, the field of medical diagnosis has witnessed remarkable advancements through the integration of data mining techniques. Predictive data mining, in particular, has emerged as a powerful tool to analyze large volumes of patient data and extract valuable insights for accurate diagnosis and prognosis. By leveraging machine learning algorithms and statistical models, predictive data mining offers a systematic approach to identify hidden patterns, trends, and relationships within medical datasets. This paper aims to propose an algorithm for a predictive data mining approach in medical diagnosis, focusing on the theoretical foundations and principles that underpin this innovative methodology. Through the utilization of this algorithm, healthcare professionals can enhance their diagnostic capabilities, facilitate early detection of diseases, and ultimately improve patient outcomes.

III. METHODOLOGY

Proposed System: After analyzing existing techniques, researchers have identified their advantages and limitations, which impact their effectiveness. To address these issues, a proposed system aims to overcome the restraints associated with current methods. The system aims to minimize inflexibility and time-consuming model building processes. It also introduces alternative parameters to enhance flexibility. Furthermore, the system strives to improve the accuracy of the generated results or verdicts, which have been a challenge in existing techniques. These advancements aim to optimize the working behavior of the proposed system.

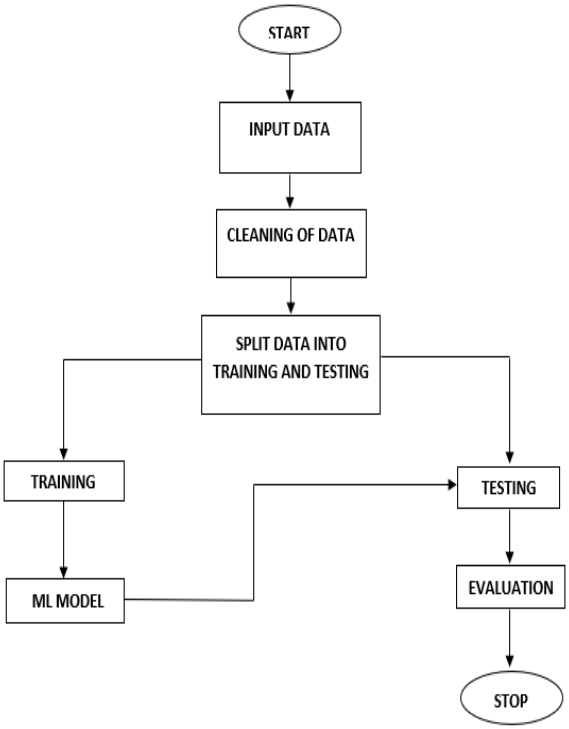

Figure 1: Block Diagram

IV. IMPLEMENTATION

The algorithms listed below were used to complete the project.

1. **Decision Tree:** A decision tree is one of the powerful and widely used machine learning techniques that predicts outcomes by creating a tree-like model of decisions and their possible consequences. It is a supervised learning method that can handle both classification and regression problems. The tree structure of internal nodes representing features, branches representing decisions, and leaf nodes representing outcomes or predictions. Decision trees are intuitive and easy to interpret, making them valuable for understanding the decision-making process. They can handle both categorical and numerical data, and handle missing values effectively. The algorithm recursively partitions the data based on the most informative features, aiming to maximize information gain or minimize impurity at each split.

   Decision trees are prone to overfitting, which can be mitigated using pruning techniques. Ensemble methods like random forests and gradient boosting can further improve their performance. Decision trees have a wide range of applications, including finance, healthcare, and marketing, where they can aid in decision-making, risk assessment, and customer segmentation. Their simplicity and interpretability make them a popular choice for various real-world problems.

2. **Random Forest:** Random Forest is a popular machine learning technique known for its versatility and robustness. It is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest is trained on a random subset of the data, and the final prediction is made by aggregating the predictions of individual trees.

   The strength of Random Forest lies in its ability to handle a variety of tasks, including classification, regression, and feature selection. It is resistant to overfitting and can handle large datasets with high-dimensional feature spaces. Random Forest also provides measures of feature importance, allowing users to identify the most influential variables in the prediction process.

   The algorithm's randomization and averaging techniques help reduce bias and variance, making it a powerful tool for both predictive accuracy and interpretability. It has been successfully applied in various fields, such as finance, healthcare, and image recognition. With its ability to handle complex problems and deliver reliable results, Random Forest continues to be a popular choice among data scientists and machine learning practitioners.

3. **Stacking Classifier:** The Stacking Classifier is an ensemble machine learning technique that combines multiple individual classifiers to create a more powerful and accurate predictive model. It leverages the concept of stacking, where the predictions of base classifiers are used as input features for a meta-classifier. The base classifiers are trained on the original dataset, while the meta-classifier is trained on the predictions made by the base classifiers. This two-level architecture allows the Stacking Classifier to capture both the individual strengths of the base classifiers and the collective decision-making power of the meta-classifier. The Stacking Classifier can be effective in scenarios where no single classifier performs exceptionally well on its own. It offers improved generalization and robustness by leveraging the diversity of base classifiers. By combining the predictions of multiple classifiers, the Stacking Classifier aims to make more accurate and reliable predictions on unseen data.

4. **XG Boost:** XGBoost, short for eXtreme Gradient Boosting, is a powerful machine learning technique that has gained immense popularity due to its effectiveness and efficiency. It belongs to the family of gradient boosting methods and is widely used in various domains, including industry and academia.

   The key idea behind XGBoost is its ability to combine multiple weak prediction models, typically decision trees, into a strong predictive model. It achieves this by iteratively building new models that focus on the errors made by the previous models. XGBoost also incorporates regularization techniques to prevent overfitting and improve generalization.

   One of the reasons for XGBoost's success is its optimization algorithm, which efficiently handles large datasets and complex models. It employs a technique called gradient boosting, where each new model is trained to minimize a differentiable loss function by considering the gradients of the loss with respect to the predictions of the previous models.

XBoost has demonstrated exceptional performance in various machine learning competitions and real-world applications, including classification, regression, ranking, and recommendation systems. Its versatility, scalability, and ability to handle diverse data types make it a valuable tool for data scientists and practitioners seeking accurate and efficient predictive modelling.

V. RESULTS AND DISCUSSION

The following results are depicted the flow and working process of project.

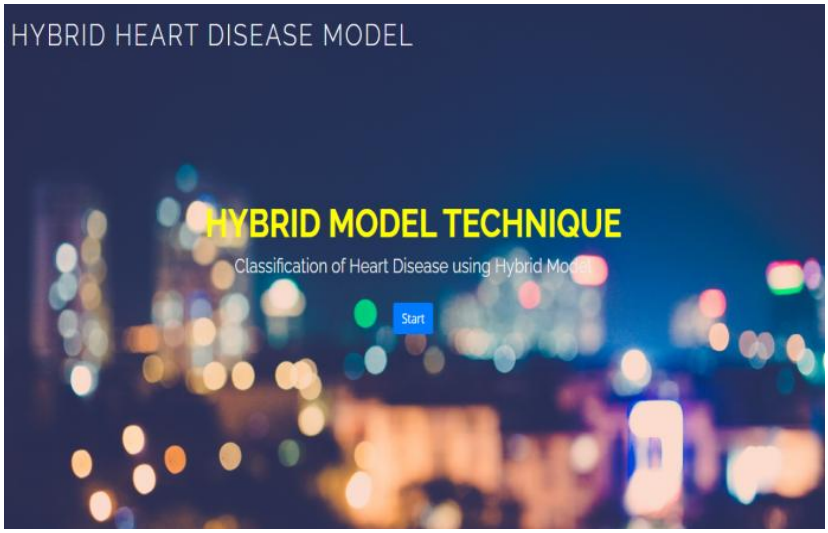

Figure 2: Home Page
Here user view the home page for Heart Conditions appellation

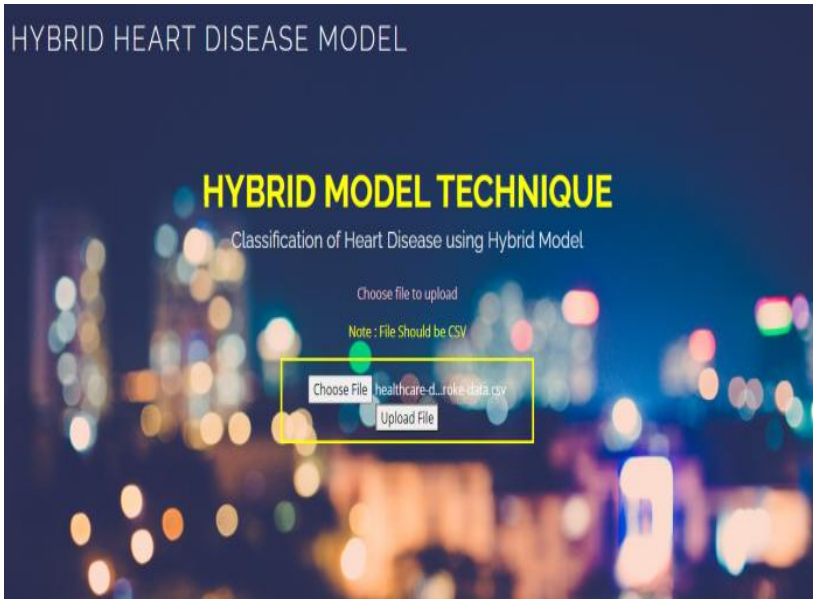

Figure 3: Load Page
User will Load the Data set

HYBRID HEART DISEASE MODEL

MODEL SELECTION

File Data

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0.0	67.0	0.0	1.0	0.0	0.0	0.0	87.96	36.6	2.0	1.0
1.0	61.0	0.0	0.0	0.0	1.0	1.0	87.96	28.1	0.0	1.0
0.0	80.0	0.0	1.0	0.0	0.0	1.0	105.92	32.5	0.0	1.0
1.0	49.0	0.0	0.0	0.0	0.0	0.0	87.96	34.4	3.0	1.0
1.0	79.0	1.0	0.0	0.0	1.0	1.0	87.96	24.0	0.0	1.0
0.0	81.0	0.0	0.0	0.0	0.0	0.0	87.96	29.0	2.0	1.0
0.0	74.0	1.0	1.0	0.0	0.0	1.0	70.09	27.4	0.0	1.0
1.0	69.0	0.0	0.0	1.0	0.0	0.0	94.39	22.8	0.0	1.0
1.0	59.0	0.0	0.0	0.0	0.0	1.0	76.15	28.1	1.0	1.0
1.0	78.0	0.0	0.0	0.0	0.0	0.0	58.57	24.2	1.0	1.0

Figure 4: View Page
User View the Data

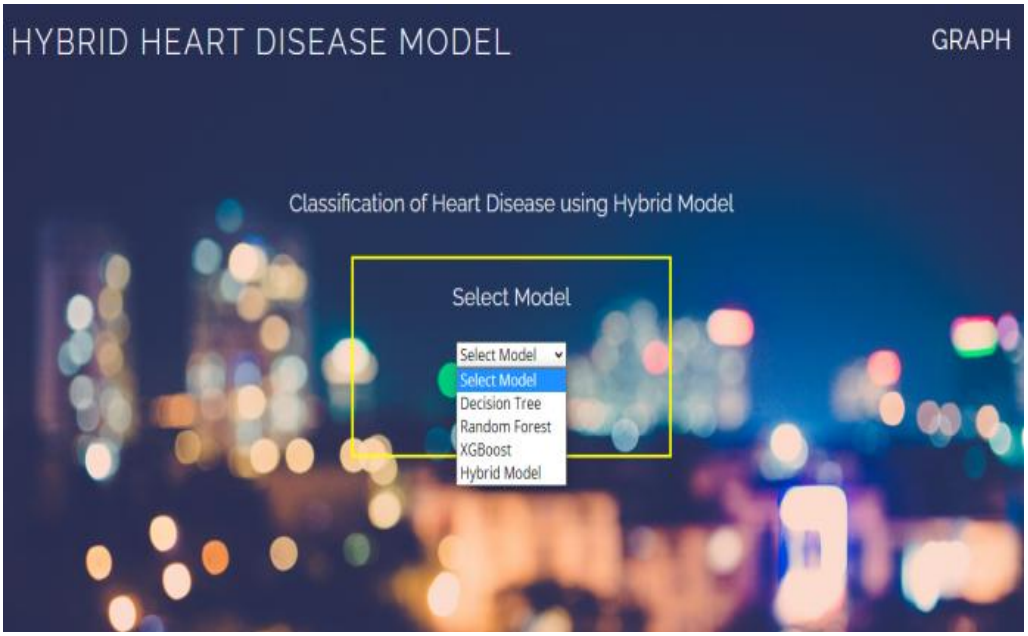

Figure 5: Model
User will View the accuracy on every algorithm

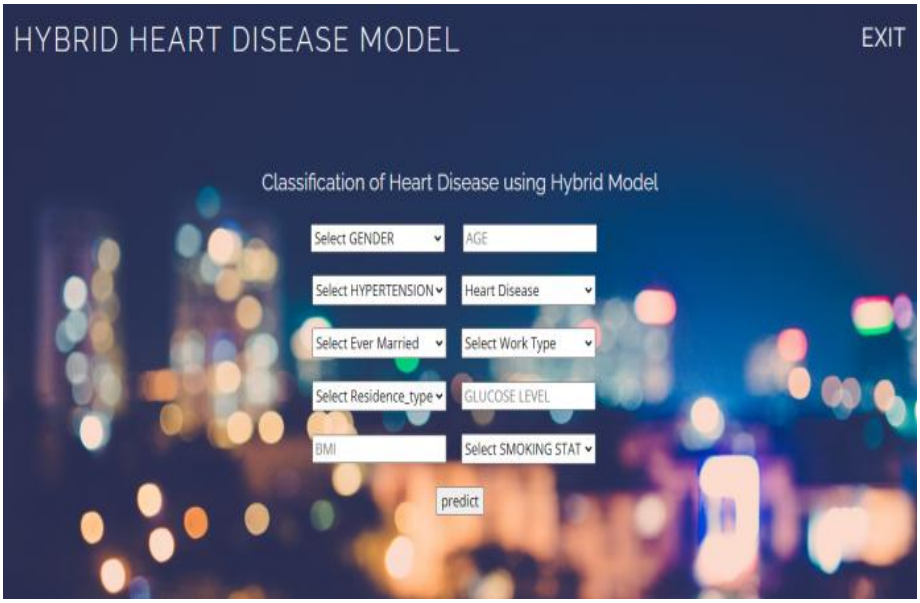

Figure 6: Prediction Page
User will give a proper input and view the result

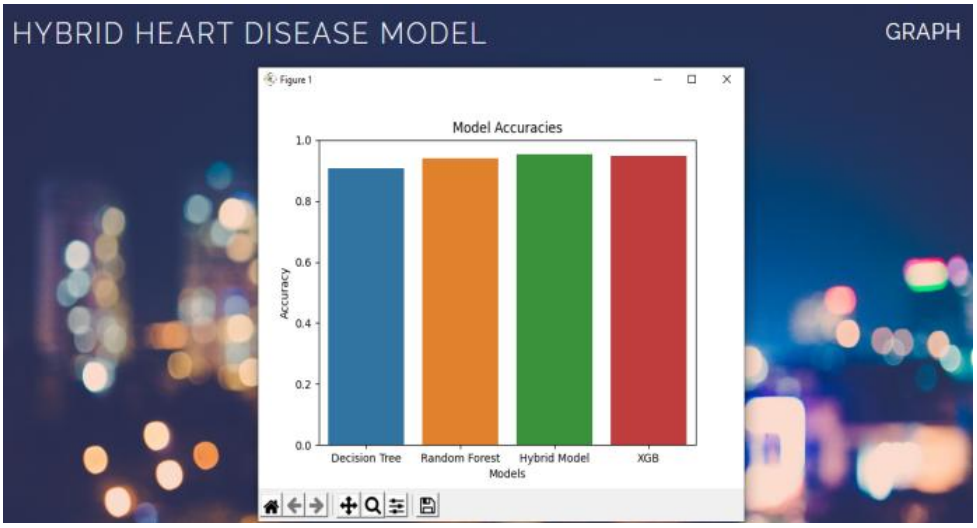

Figure 7: Graph page

VI. CONCLUSION

This investigation focused on enhancing efficiency, suitability, and Quality of Service (QoS) through the development of a more effective method. By analyzing existing methods' characteristics and limitations in the literature survey, the study aimed to overcome these shortcomings. Four algorithms, namely Random Forest, XGBoost, and two variations of Decision Tree were thoroughly investigated. The proposed method robustly evaluated these algorithms using statistical analysis and selected the most suitable pair, employing a linear model based on feature selection through best-first search, Gain ratio, and the Ranker

method. Through multiple simulations, the proposed approach consistently demonstrated its superiority, effectively addressing the limitations of both traditional and modern algorithms.

REFERENCES

[1] Ritu. Sharma, Mr Shiv Kumar, Mr. RohitMaheshwari "Comparative Analysis of Classification Techniques in DataMining Using Different Datasets" International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 4, Issue. 12, December 2015, pp.-125 – 134.

[2] SobhanSarkar, Atul Patel, SarthakMadaan, JhareswarMaiti "Prediction of Occupational Accidents Using DecisionTree Approach" IEEE Annual India Conference (INDICON), 2016, pp.- 1-6.

[3] AayushiVerma, Shikha Mehta "A Comparative Study of Ensemble LearningMethods for Classification in Bioinformatics" IEEE 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence, 2017, pp.- 155-158.

[4] K. C. Giri, M. Patel, A. Sinhal and D. Gautam, "A Novel Paradigm of Melanoma Diagnosis Using Machine Learning and Information Theory," 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE), Sathyamangalam, Tamil Nadu, India, 2019, pp. 1-7, doi: 10.1109/ICACCE46606.2019.9079975.

[5] AyisheshimAlmaw, KalyaniKadam "Survey Paper on Crime Prediction using EnsembleApproach" International Journal of Pure and Applied Mathematics, Volume 118 No. 8 2018, pp.-133-139.

[6] ShakuntalaJatav and Vivek Sharma "An Algorithm For Predictive DataMining Approach In Medical Diagnosis" International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018, pp.- 11-20.

[7] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang "Type 2 diabetes mellitus prediction model based on data mining" ELSEVIER Informatics in Medicine Unlocked, 2018, pp.- 100-107.

[8] Patel M., Choudhary N. (2017) Designing an Enhanced Simulation Module for Multimedia Transmission Over Wireless Standards. In: Modi N., Verma P., Trivedi B. (eds) Proceedings of International Conference on

[9] Communication and Networks. Advances in Intelligent Systems and Computing, vol 508. Springer, Singapore. https://doi.org/10.1007/978-981-10-2750-

[10] Sumalatha.V , Dr.Santhi.R "A Study on Hidden Markov Model (HMM)" International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 11, November 2014, pp.- 465-469.

[11] Zhang Youzhi "Research and Application of Hidden Markov Model in Data Mining" Second IITA International Conference on Geoscience and Remote Sensing, IEEE, 2010, pp.-459-462.

[12] PadmavathiJanardhanan, Heena L., and FathimaSabika "Effectiveness Of Support Vector Machines In Medical Data Mining" Journal Of Communications Software And Systems, Vol. 11, No. 1, March 2015, pp.- 25-30.

[13] GaganjotKaur, AmitChhabra "Improved J48 Classification Algorithm for the Prediction of Diabetes" International Journal of Computer Applications, Volume 98 – No.22, July 2014, pp.- 13-17.

[14] Senthilkumar Mohan, ChandrasegarThirumalai, &GautamSrivastava "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" Special Section On Smart Caching, Communications, Computing And Cybersecurity For Information-Centric Internet of Things, IEEE, 2019, pp.- 81542-81554.

[15] Shekhawat V.S., Tiwari M., Patel M. (2021) A Secured Steganography Algorithm for Hiding an Image and Data in an Image Using LSB Technique. In: Singh V., Asari V.K., Kumar S., Patel R.B. (eds) Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing, vol 1257. Springer, Singapore. https://doi.org/10.1007/978-981-15-7907-3_35.

[16] H. Gupta and M. Patel, "Study of Extractive Text Summarizer Using The Elmo Embedding," 2020 Fourth International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 829-834, doi: 10.1109/ISMAC49090.2020.9243610

