ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering &Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Text Summarization with Heart Disease Prediction

Mrs. Preethi M[1], Wilton Lobo[2], Shravya[3], Suraksha S Salian[4], Prajna R Shetty[5]

Asst. Professor, Department of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India [1]

Student, Department of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India [2,3,4,5]

ABSTRACT: Text summarization and heart disease prediction are two important areas of research that have the potential to significantly impact society. Text summarization involves generating a brief and coherent summary of a document while retaining its most important information. Heart disease prediction involves using machine learning algorithms to identify individuals who are at a higher risk of developing cardiovascular disease. There are various techniques and algorithms for text summarization, including extraction-based, abstraction-based, and hybrid methods. These methods can be useful in fields such as medicine, law, and journalism, where it is becoming increasingly difficult to quickly understand and extract the most important information from a document. Based on a variety of risk indicators, including age, gender, blood pressure, cholesterol levels, and smoking status, machine learning algorithms have been used to estimate a person's likelihood of getting heart disease. Heart disease prevention and early detection are crucial areas of research since they can greatly improve patient outcomes and lower healthcare expenditures. However, accuracy and bias problems exist for both text summarizing and heart disease prediction. In text summarization, accuracy and brevity must be balanced, whereas biases in the selection of risk variables and the creation of machine learning algorithms can affect the prediction of heart disease. To identify those who are more likely to develop cardiovascular disease, many techniques and models are used for heart disease prediction.

I. INTRODUCTION

Text summary is a strategy for condensing a text while keeping the important details. The analysis of data and prediction of future coronary heart disease (CHD) risk using machine learning algorithms and deep learning is a significant field of research. Machine learning algorithms including classification analysis, regression, data clustering, feature engineering, and dimensionality reduction are used to create efficient data mining strategies that can accurately forecast the onset of cardiac disease. The outputs of these models are also interpreted using explainable AI methods. The UCI Machine Learning Repository is one of the largest datasets for this application. When there is a lack of data, artificial data can be utilized to address privacy issues and

II. LITERATURE REVIEW

Text summarization has been applied to the field of heart disease prediction to assist medical practitioners in decision-making processes. Studies have utilized various machine learning techniques such as recurrent neural networks, support vector machines, and deep learning models to summarize medical texts and predict heart disease risk. The evaluation of these models has produced encouraging outcomes in terms of performance and accuracy. However, further research is needed to improve the generalization and interpretability of these models for practical use in healthcare settings.

Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava proposed a Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques [1]: The proposed approach combines decision trees, logistic regression, and artificial neural networks to generate a comprehensive model for predicting heart disease risk. The authors conducted experiments on real-world datasets and achieved high levels of accuracy in heart disease prediction. They compared the performance of the hybrid model with several traditional machine learning algorithms and found that the hybrid model outperformed other models in terms of accuracy, sensitivity, and specificity. The study demonstrates the potential of combining multiple machine learning techniques to improve the accuracy and performance of heart disease prediction models. The hybrid model can effectively leverage the strengths of different machine learning algorithms and provide more accurate and reliable predictions for heart disease risk. The study

highlights the importance of developing innovative machine learning approaches for effective heart disease prediction and management.

V.V.Ramalingam proposed Heart Disease Prediction Using Machine Learning Algorithms [2]: Key challenges in heart disease prediction, including data imbalance, missing values, and feature selection, and discusses how machine learning algorithms can address these challenges. The review covers various machine learning algorithms, including decision trees, logistic regression, support vector machines, and artificial neural networks. The author provides a detailed overview of each algorithm, including their strengths and weaknesses, and highlights their applications in heart disease prediction. Also discusses the importance of data pre-processing and feature selection in heart disease prediction and provides insights into the best practices for data pre-processing and feature selection.

Avinash Golande, Pavan Kumar T proposed Heart Disease Prediction Using Effective Machine Learning Techniques [3]: The importance of heart disease prediction in clinical settings and the challenges associated with predicting heart disease risk accurately. The review covers various machine learning techniques, including decision trees, logistic regression, support vector machines, and artificial neural networks. The authors provide a detailed overview of each technique, including their strengths and weaknesses, and highlight their applications in heart disease prediction. It also discusses the importance of data pre-processing, feature selection, and model evaluation in heart disease prediction and provide insights into the best practices for each step of the machine learning pipeline. The paper also identifies areas for future research and development, including the use of deep learning models and the integration of multiple data sources for heart disease prediction.

Mr.Santhana Krishnan.J, Dr.Geetha.S proposed Prediction of Heart Disease Using Machine Learning Algorithms [4]: Discusses the importance of data pre-processing, feature selection, and model evaluation in heart disease prediction and provide insights into the best practices for each step of the machine learning pipeline. The paper provides a valuable resource for researchers and practitioners in the field of heart disease prediction. The review highlights the potential of machine learning algorithms in addressing the challenges of heart disease prediction and provides guidance on selecting the most appropriate algorithm for a given dataset. The paper also identifies areas for future research and development, including the use of ensemble methods and the integration of multimodal data for heart disease prediction.

Archana Singh and Rakesh Kumar proposed heart disease prediction using machine learning algorithms [5]. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence (AI), it provides prestigious support in predicting any kind of event which take training from natural events. In this paper, we calculate accuracy of machine learning algorithms for predicting heart disease, for this algorithm are k-nearest neighbor, decision tree, linear regression and support vector machine (SVM) by using UCI repository dataset for training and testing.

No.	Paper Title	Author Name	Key Points	Remark
1	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava 2019	1. Novel method that aims at finding significant features by applying machine learning techniques resulting in the prediction of cardiovascular disease. 2. The prediction model is introduced with different combinations of features and several known classification techniques such as Random Forest, Linear Model.	Produces an enhanced performance level through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).
2	Heart Disease Prediction Using Machine Learning Algorithms	V.V.Ramalingam 2018	1)Paper evaluates and analyse the performance of numerous models such as, Decision Trees, K Nearest Neighbor, Naive Bayes, and SVM based on such methods and methodologies.	SVM has more accuracy than other techniques.

			2) Researchers favor models based on supervised learning techniques including SVM, KNN, Naive Bayes, DT, Random Forest (RF), and ensemble models.	
3	Heart Disease Prediction Using Effective Machine Learning Techniques	Avinash Golande, Pavan Kumar T 2019	Paper presents a survey of various models based on such algorithms and techniques and analyze their performance.	Models based on supervised learning algorithms produces an enhanced performance.
4	Prediction of Heart Disease Using Machine Learning Algorithms	Mr.Santhana Krishnan.J, Dr.Geetha.S 2019	Aim of this system is to predict the possibilities of occurring heart disease of the patients in terms of percentage. This is performed through data mining classification techniques.	System predicts the arising possibilities of heart disease.
5	Heart disease prediction using machine learning algorithms	Archana Singh, and Rakesh Kumar, 2020.	Paper includes investigation on several machine learning algorithms to predict cardiac disorders using the dataset with 14 features of the patient from the UCI ML repository.	Performance of the Random Forest algorithm is significantly better.

The research builds on earlier studies that have explored the impact of these factors on heart disease. The main idea is to be able to find a short subset of the most essential information from the entire set and present it in a human-readable format. As online textual data grows, automatic text summarization methods have potential to be very helpful because more useful information can be read in a short time. By examining these factors from a holistic perspective, the authors hope to provide insights that can inform the design of more effective interface for prediction of heart disease.

III. METHODOLOGY OF PROPOSED SURVEY Graphical User Interface (GUI):

Healthcare practitioners that could lack programming or data science competence should be taken into consideration when creating the GUI. It must be simple to use, straightforward, and offer accurate results visualizations. Additionally, the GUI should make it simple for users to enter patient data and choose pertinent aspects for analysis. Using APIs or web services, machine learning models can be included into the GUI. Overall, GUIs are essential in enabling healthcare workers to use machine learning models even if they lack programming or data science skills. By giving precise projections of future coronary heart disease (CHD) risk based on patient data, they help healthcare practitioners make knowledgeable decisions about patient care.

ALGORITHMS:

SVM Classifier:

Popular supervised machine learning approach called Support Vector Machine (SVM) is used to solve classification and regression issues. Finding a hyperplane in an N-dimensional space that divides the data into distinct classes is the goal of the SVM method. Finding a maximum-margin hyperplane that splits the dataset into two classes most effectively is the foundation of SVMs. The gap between the nearest data points from each class and the hyperplane is referred to as the margin. This margin is maximized by the SVM method, which improves generalization performance. SVMs are frequently utilised in many different areas, including bioinformatics, text classification, and picture classification. When dealing with high-dimensional datasets, when other techniques might not work effectively, they are especially helpful. By transforming the input data into a higher-dimensional space where it may be linearly separated using kernel functions, SVMs can even handle data that is not linearly separable. SVMs are robust machine learning algorithms that can be applied to both classification and regression issues in general. They are frequently utilised in many applications and have been demonstrated to work effectively on a variety of datasets. Data points are mapped into higher dimensional spaces by the SVM algorithm in order to facilitate easier classification. Kernel functions like linear, polynomial, radial basis function (RBF), or sigmoid kernels are used for this.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Random Forest Classifier:

Using bootstrapped samples of the training data, the random forest algorithm builds several decision trees. Overfitting is lessened by the fact that each tree is trained using a separate collection of features. Using either majority voting (for classification) or averaging (for regression), the final forecast is formed by combining the predictions of all the trees in the forest. In general, random forests are effective machine learning algorithms that may be applied to a variety of tasks. They can effectively manage missing or noisy data, need little feature engineering, and are simple to use. A lot of decision trees are created using the Random Forest Classifier, and each one is trained using a different random subset of the data. This lessens overfitting, a problem that decision trees frequently experience. Based on its own set of rules, each decision tree in the forest determines the class of a given input. By considering the consensus of all the decision trees in the forest, the final prediction is made. Both binary and multiclass classification jobs can employ the technique. The final prediction is made from the class with the highest likelihood, which is represented in the Random Forest Classifier's output as a probability distribution over all conceivable classes.

There are a number of hyperparameters that can be altered to enhance the performance of the Random Forest Classifier. The number of trees in the forest, each tree's maximum depth, and the number of features utilised to divide each node are the three most crucial hyperparameters. The accuracy of the model can be increased and overfitting can be minimised by tuning these hyperparameters.The dataset is randomly divided into training and testing sets in order to train a Random Forest Classifier. Following training on the training set, the performance of the decision trees is assessed on the testing set. To make sure the model is adaptable to various training and testing sets, this procedure is performed numerous times.

KNeighbors Classifier:

Calculating the distance between a query and each of the instances in the educational data is the core concept of the KNeighborsClassifier. The gap could, for instance, be the vectors' Euclidean distance from one another. The space is determined, and then you select the best k instances that are most closely connected to the question. A few number of straightforward algorithms, such linear regression or support vector machines, are particularly effective in solving the majority of logical device learning problems. However, there are some situations where more complex algorithms are beneficial. In particular, the KNeighborsClassifier is a wonderful choice if you require a non-linear model and your statistics are not well characterised by linear models. KNN is a straightforward algorithm and non-parametric method that may be applied to any kind of regression. KNN can be applied to data types with more training data. By using the average of the good nearest neighbours as the estimate of the based variable for a certain independent variable, KNN can be utilised for regression.

Methodology

The summarizer is designed using the algorithm called DistilBert. It works by filtering the words that are only related to the heart diseases. We provide a way for pre-training the DistilBERT general-purpose language representation model, which can be customised to perform well on a variety of tasks like its larger cousin:

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

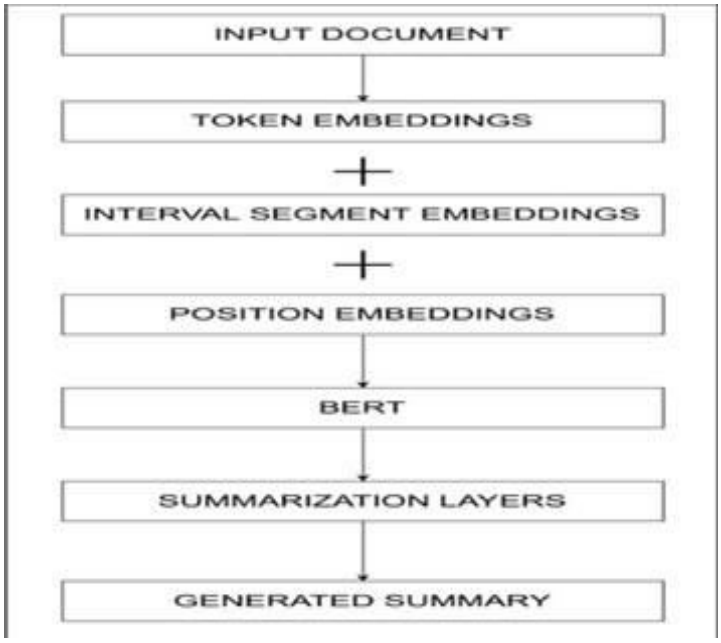

Fig -1: Methodology of Text Summarizer

Data Source:

Clinical databases provide a substantial amount of information on individuals and their medical conditions. The term "heart diseases" refers to a broad spectrum of conditions that affect the heart. The leading cause of death in the world is cardiovascular disease. Various conditions that affect the heart, blood arteries, and the body's capacity to pump and circulate blood are referred to as "cardiovascular disease" as a whole. Records with medical attributes were provided through the Kaggle heart disease database.

Pre-processing Data:

After records have been gathered, the facts are uniformized and the information is processed. Item-kind functions that needed to be translated into a numbers datatype are present in the retrieved dataset. An important step in text summarization and heart disease prediction is pre-processing the data. Pre-processing in text summarization entails purging the text data of stop words, punctuation, and other superfluous information. Following the tokenization of the text into individual words or phrases, different summarization approaches, such as extractive or abstractive summarization, are employed to create the final summary. Pre-processing in heart disease prediction entails cleaning up and converting patient data into a format that machine learning algorithms can exploit. This could entail picking pertinent features for analysis, addressing missing values, and normalising the data.In both situations, pre-processing enhances model accuracy by removing noise and unimportant variables from the input. Additionally, it ensures that the models are reliable and effectively generalise to new data. For pre-processing text and patient data, there are numerous tools and libraries available. A well-liked library for natural language processing tasks including tokenization, stemming, and lemmatization is NLTK (Natural Language Toolkit). Another well-known package that offers numerous techniques for pre-processing numerical data, such as scaling and normalisation, is Scikit-learn.Pre-processing is crucial for text summarization and heart disease prediction overall. It assists in ensuring that the input data is accurate, pertinent, and in a format that machine learning algorithms can use efficiently.

Design of Random Forest, Support Vector Machine (SVM) and KNeighbors Classifier:

Popular machine learning techniques for classification and regression issues include Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). In order to apply these algorithms, one must first get the data ready by cleaning, pre-processing, and converting it into a format that the algorithm can use. This could entail picking pertinent features for analysis, addressing missing values, and normalising the data. Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) are popular machine learning approaches for classification and regression problems. Cleaning, pre-processing, and transforming the data into a format that the algorithm can use are

necessary before applying these methods. Choosing the most relevant features for analysis, dealing with missing values, and normalising the data may all be necessary.

Testing and Classifying Output:

After the intended neural community has been trained, Support vector machines, Random Forests, and KNeighborsClassifiers are tested using test data. After the classification accuracy of the classifier is assessed using test data, the most accurate approach is employed to forecast or deliver high-quality results. Random Forest, an ensemble learning technique used in this challenge, blends various decision trees to increase prediction accuracy. When working with high-dimensional datasets, when other techniques might not perform well, it is especially helpful. You may gauge Random Forest's effectiveness by examining the precision of its forecasts on a validation dataset.

IV. CONCLUSION AND FUTURE WORK

These findings demonstrate that, despite the fact that the majority of studies diagnose heart disease using classifier methods like Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Random Forest. The software has the potential to completely change how medical professionals examine patient records and estimate their risk of developing heart disease. The research intends to increase the precision and timeliness of heart disease diagnosis by using natural language processing and machine learning techniques to extract crucial information from medical records and identify risk factors for heart disease. Both patients and healthcare providers stand to gain significantly from the programme, which draws on previous research in the areas of medical record analysis and heart disease prediction. Healthcare professionals can create individualised treatment plans and preventative actions that can help lower the incidence of heart disease and improve patient outcomes by precisely estimating the risk of heart disease in individuals.

In order to increase the precision of the predictive models, future work for the Text Summarization with Heart Disease Prediction project may involve incorporating more data sources, such as wearable technology and mobile health applications. AI strategies that are easy to understand could produce more transparent and understandable findings, increasing patient confidence and acceptance in healthcare settings. The usefulness and efficacy of the predictive models might be increased by creating a user-friendly interface for healthcare practitioners, which would ultimately result in better patient outcomes and higher levels of healthcare quality.

REFERENCES

1. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE, 19 June 2019.
2. V.V. Ramalingam “Heart disease prediction using machine learning techniques”, International Journal of Engineering & Technology, March 2018.
3. Avinash Golande, Pavan Kumar T “Heart Disease Prediction Using Effective Machine Learning Techniques” International Journal of Recent Technology and Engineering (IJRTE), June 2019.
4. Mr.Santhana Krishnan.J, Dr.Geetha.S “Prediction of Heart Disease Using Machine Learning Algorithms” IEEE, 2019
5. Archana Singh and Rakesh Kumar “Heart Disease Prediction Using Machine Learning Algorithms” 2020 International Conference on Electrical and Electronics Engineering (ICE3), February 2020.
6. Dwivedi, Ashok Kumar. "Performance evaluation of different machine learning techniques for prediction of heart disease." Neural Computing and Applications 29, no. 10 (2018): 685-693.
7. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc.Int.Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012,pp. 22–25.
8. Prediction of Heart Diseases using Random Forest.To cite this article: Madhumita Pal and Smita Parija 2021 J. Phys.: Conf. Ser. 1817 012009.

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com