Disease prediction Model using Machine Learning

Saswat Bhotto Mishra
School of Electronics Engineering
Vellore Institute of technology
Vellore, India
saswatbhotto.mishra2020@vitstudent.ac.in

Malay Rajpoot
School of Electronics Engineering
Vellore Institute of technology
Vellore, India
malay.rajpoot2020@vitstudent.ac.in

Veerayya Vastrad
School of Electronics Engineering
Vellore Institute of technology
Vellore, India
veerayya.vastrad2020@vitstudent.ac.in

Utkaarsh Dubey
School of Electronics Engineering
Vellore Institute of technology
Vellore, India
utkaarshdubey.s2020@vitstudent.ac.in

Dr. Aparna Mohanty
School of Electronics Engineering
Vellore Institute of technology
Vellore, India
aparna.mohanty@vit.ac.in

***Abstract*—People today deal with a variety of ailments as a result of their lifestyle choices and the surroundings. Therefore, it becomes crucial to make disease predictions at an earlier stage. But doctors find it too challenging to make an accurate forecast based on symptoms. The hardest task is making an accurate diagnosis of a condition. Data mining is crucial in predicting the sickness in order to solve this issue. Each year, there is significant data increase in the medical sciences. The accurate analysis of medical data that has benefited from early patient care has grown as the amount of data in the medical and healthcare fields has increased. Data mining identifies hidden pattern information ina vast amount of medical data with the use of disease data.**

***Keywords*—Data Preparation or mining (Kaggle), Cleaning, Support Vector Classifier, Naive Bayes Classifier, and Random Forest Classifier (Model Building)**

I. Introduction

Making educated judgements about patient care and treatment can help medical personnel, thanks to the development of illness prediction models utilizing machine learning algorithms. These models can examine enormous volumes of data and spot patterns that aren't always visible to the naked eye, giving us insights into how diseases develop and allowing us to identify high-risk patients who can benefit from early intervention. Machine learning algorithms can also be used to forecast the course of an illness, including mortality, recurrence, and therapy response rates. As a result, clinicians may be able to develop individualized treatment programs for each patient, improving both their health and quality of life. With the use of machine learning methods, we intendto create a disease prediction model in this project. Publicly accessible medical data will be used by us to diseases. In this project, we want to use machine learning to create an illness prediction model. To train and evaluate our model, we will use medical data that is openly accessible. Different machine learning algorithms will be used to build the model. Identify applicable funding agency here. If none, delete this, and the effectiveness of each algorithm will be compared in order to choose the best strategy. The ultimate objective of thisresearch is to create a disease prediction model that is precise and trustworthy and that can be applied in clinical practice to enhance patient care and outcomes

II. Motivation

Designing a disease prediction model using machine learning techniques is driven by the need to enhance healthcare outcomes by giving doctors precise and timely patient health information. We put forth a general disease prediction based on the patient's symptoms. We will train the Support Vector Classifier, Naive Bayes Classifier, and Random Forest Classifier using the collected data in order to forecast diseases. To assess the models' quality, a confusion matrix will be used. Once the three models have been trained, we will combine their predictions to forecast the disease from the input symptoms. This strengthens and improves the accuracy of our total prediction. The function will predict the disease based on the symptoms using the trained models, deliver the predictions in a JSON format, and accept symptoms separated by commas as input. For improved health outcomes and lower healthcare costs, early disease detection and treatment are crucial. Traditional diagnostic techniques, however, can sometimes be costly, time- consuming, and unreliable. Electronic health records, medical pictures, and patient demographics can all be analyzed bymachine learning algorithms to find trends that would not be visible to a human doctor. Additionally, applying machine learning algorithms to disease prediction models can assist medical professionals in making better choices regarding patient care and treatment. Doctors can intervene early and stop the beginning or progression of diseases by identifying high- risk patients. Additionally, illness prediction models can give medical professionals insights into how patients will respond to treatment, allowing them to design individualized plans of care that are specific to each patient's need. Overall, by

increasing the precision and effectiveness of disease diagnosis and prediction, the creation of disease prediction models utilizing machine learning techniques has the potential to revolutionize healthcare. By creating a precise and trustworthy illness prediction model that can be applied in clinical practice to enhance patient outcomes, our research intends to make a contribution to this expanding field. creating a reliable and accurate illness prediction model for clinical usage to enhance patient outcomes

III. LITERATURE REVIEW

M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, in 2017 , in [1] "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, no. 1, pp. 8869-8879, 2017. The paper "Disease prediction by machine learning over big data from healthcare communities" by Chen et al. aims to predict diseases using machine learning techniques on big data obtained from healthcare communities. The results of the study show that machine learningmod- els can accurately predict diseases using big data from healthcare communities. The authors note that logistic regression and support vector machine models performed the best, with accuracies above 90 percent. They also found that demographic and social media data can be useful predictors of disease, in addition to traditional medical data.The authors argue that utilizing the massive amount of data generated from healthcare communities can help identify early signs of diseases, enabling early diagnosis and treatment. The paper provides valuable insights into the potential of machine learning techniques to aid in diseaseprediction using big data from healthcare communities. The study's findings have important implications for healthcare professionals, policymakers, and researchers interested in improving early diagnosis and treatment of diseases. However, the paper is limited by its focus on a single dataset, and future research should explore the generalizability of these findings to other healthcare communities and populations.

B. Qian, X. Wang, N. Cao, H. Li and Y.-G. Jiang, in [2] have made a paper on "A relative similarity based method for interactive patient risk prediction", Springer Data Mining Knowledge . Discovery, vol. 29, no. 4, pp. 1070-1093, 2015. The paper "A relative similarity based method for interactive patient risk prediction" by Qian et al. (2015) presents a new method for interactive patient risk prediction that uses relative similarity measures. The authors argue that traditional risk prediction models are limited by their inability to handle complex patient data, and that their method can overcome these limitations by incorporating multiple sources of data and allowing for user interaction. , The paper provides a valuable contribution to the field of patient risk prediction by proposing a new methodthat can handle complex patient data and incorporate user interaction. The study's findings have important implications for healthcare professionals and policymakers interested in improving the accuracy of patient risk predictions. However, the paper is limited by its focuson a single dataset, and future research should explore the generalizability of these findings to other patient populations and healthcare settings.

IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang and C. Youn, in [3] have researched on, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system", IEEE Communication, vol. 55, no. 1, pp. 54-61, Jan. 2017. The paper "Wearable 2.0: Enable human-cloud integration in next generation healthcare system" by Chen et al. (2017) discusses the potential of wearable technology to enable human-cloud integration in the next generation of healthcare systems. The authors argue that wearable technology can help improve the quality of care, reduce healthcare costs, and enhance patient engagement. The authors also discuss the challenges of implementing Wearable 2.0 in healthcare, including privacy concerns, data security, and interoperability issues. They propose several solutions to address these challenges, such as using secure cloud-based platforms and implementing standardized protocols for data exchange. The paper also presents several use cases of Wearable2.0 in healthcare, including remote monitoring of chronic diseases, personalized treatment plans, and early detection of health issues. The authors highlight the potential benefits of these use cases, including improved patient outcomes, reduced healthcare costs, and enhanced patient engagement. The paper provides a comprehensive overview of the potential of Wearable 2.0 to enable human-cloud integration in the next generation of healthcare systems. The authors' proposed solutions for addressing the challenges of implementing wearable technology in healthcare are with the widespread adoption of Wearable 2.0 in healthcare.

Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan and A. Alamri, in [4] made a paper "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data", IEEE Syst. J, vol. 11,no. 1, pp. 88-95, Mar. 2017.The paper "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data" by Zhang et al. (2017) proposes a healthcare cyberphysical system (CPS) that leverages cloud computing and big data analytics to improve the quality of care and reduce healthcare costs. The paper provides a valuable contribution to the field of healthcare systems by proposing a healthcare CPS that leverages cloud computing and big data analytics.

The authors also discuss the challenges of implementing HealthCPS, including data privacy and security concerns,

interoperability issues, and the need for standardized protocols. They propose several solutions to address these challenges, such as using secure cloud-based platforms and implementing standardized data exchange protocols. The paper also presents several use cases of HealthCPS in healthcare, including remote monitoring of chronic diseases, real-time analysis of patient data for early detection of health issues, and personalized treatment plans based on patient data. The authors highlight the potential benefits of these use cases, including improved patient outcomes, reduced healthcare costs, and enhanced patient engagement.The authors' proposed solutions for addressing the challenges of implementing HealthCPS are also valuable contributions to the field. However, the paper is limited by its focus on the potential benefits of HealthCPS, and future research should also explore potential risks and ethical concerns associated with the widespread adoption of HealthCPS in healthcare.

L. Qiu, K. Gai and M. Qiu, in [5] researched on" Optimal big data sharing approach for telehealth in cloud computing", Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), pp. 184-189, Nov.2016.The paper "Optimal big data sharing approach for tele- health in cloud computing" by Qiu et al. (2016) proposes an optimal approach for big data sharing in telehealth using cloud computing. The authors argue that telehealth can improve access to healthcare services, reduce health- care costs, and enhance patient engagement. However, the use of telehealth generates a large amount of datathat must be securely and efficiently managed. Overall, the paper provides a valuable contribution to thefield of telehealth by proposing an optimal approach forbig data sharing using cloud computing. The authors'proposed hierarchical structure for data sharing and dataencryption scheme are valuable contributions to the field. However, the paper is limited by its focus on the technicalaspects of data sharing and future research should alsoexplore the potential impact of the proposed approach onpatient outcomes and healthcare costs.

Ajinkya Kunjir, Harshal Sawant and Nuzhat F. Shaikh, made [6], a paper on "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare", IEEE big data analytics and computational intelligence, pp. 2325, Oct 2017. The paper "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare" by Kunjir et al. (2017) proposes a data mining and visualization approach for predicting multiple diseases in healthcare. The authors argue that early detection of diseases can lead to better patient outcomes and reduced healthcare costs. However, the process of identifying multiple diseases can be challenging due to the large amount of patient data that must be analyzed. Overall, the paper provides a valuable contribution to the field of healthcare systems by proposing a data mining and visualization approach for predicting multiple diseases. The authors' proposed two-stage classification approach and visualization approach are valuable contri- butions to the field. However, the paper is limited by its focus on a single case study, and future research should also explore the potential generalizability of the proposed approach to other healthcare settings.

Shanthi Mendis, Pekka Puska and Bo Norrving, in [7] made "Global Atlas on Cardiovascular Disease Prevention andControl",pp. 3-18, 2011, ISBN 978-92-4-156437-3. The paper "Global Atlas on Cardiovascular Disease Pre- vention and Control" by Mendis et al. (2011) is acompre- hensive overview of the global burden of cardiovascular disease (CVD) and strategies for prevention and control. The authors argue that CVD is amajor public health concern worldwide, and the burden of the disease is expected to increase in the future due to aging populations and changes in lifestyle factors. , the paper provides a valuable contribution tothe field of global health by highlighting the global burden of CVD and proposing strategies for prevention and control. The authors' comprehensive approach to addressing the social and economic determinants of CVDand the importance of a multi-sectoral approach are valuable contributions to the field. However, the paper is limited by its focus on policy and intervention strategies, and future research should also explore the effectiveness of these strategies in different settings.

In [8] , S.U. Amin, K. Agarwal and R. Beg, "Genetic neuralnetwork based data mining in prediction of heart disease using risk factors", IEEE Conference on Information Communication Technologies (ICT), pp. 1227-31, 11-12 April 2013.The paper "Genetic neural network based data mining in prediction of heart disease using risk factors" by Amin et al. (2013) proposes a novel approach to predicting heart disease using a genetic neural network (GNN) and risk factors. The authors argue that heart disease is a major cause of mortality worldwide, and accurate prediction of the disease can help in early detection and

management. The authors tested their approach using a dataset of 303 patients with 14 risk factors, including age, sex, blood pressure, and cholesterol levels. The authors trained and tested the GNN on the dataset and achieved a prediction accuracy of 91.08 percent, which is higher than other approaches such as decision trees and logistic regression. The authors argue that the GNN approach has several advantages, including the ability to handle non-linear relationships between risk factors and the outcome, as well as the ability to automatically select the most relevant risk factors for prediction. The authors also suggest that their approach can be applied to other medical prediction tasks beyond heart disease. The paper provides a valuable contribution to the field of data mining and healthcare by proposing a novel approach to predicting heart disease using a GNN and risk factors. The authors' focus on optimizing predictive performance and the ability to handle non- linear relationships are valuable contributions to the field. However, the paper is limited by its focus on a single dataset and the absence of external validation of the proposed approach. Future research should explore the effectiveness of the GNN approach in other settings and populations.

In [9] , S Palaniappan and R Awang, "Intelligent heart dis- ease prediction System using data mining techniques", IEEE/ACS International Conference on Computer Systems and Applications AICCSA 2008, pp. 108115, March31 2008-April 4 2008. The research paper "Intelligent heart disease prediction system using data mining techniques" by Palaniappan and Awang (2008) proposes a novel approach to predicting heart disease using data mining techniques. The authors argue that heart disease is a leading cause of death worldwide, and early detection can significantly reduce morbidity and mortality. The authors tested their approach using a dataset of 920 patients with 15 risk factors, including age, sex, blood pressure, and cholesterol levels. The authors trained and tested the models using the dataset and achieved a prediction accuracy of up to 90percent , which is higher than other approaches such as logistic regression. The authors argue that their approach has several advantages, including the ability to handle non-linear relation- ships between risk factors and the outcome, as well as the ability to automatically select the most relevant risk factors for prediction. Overall, the paper provides a valuable contribution to the field of data mining and healthcare by proposing a novel approach to predicting heart disease using data mining techniques. The authors' focus on optimizingpredictive performance and the ability to handle non- linear relationships are valuable contributions to the field. However, the paper is limited by its focus on a single dataset and the absence of external validation of the proposed approach. Future research should explore the effectiveness of the approach in other settings and populations.

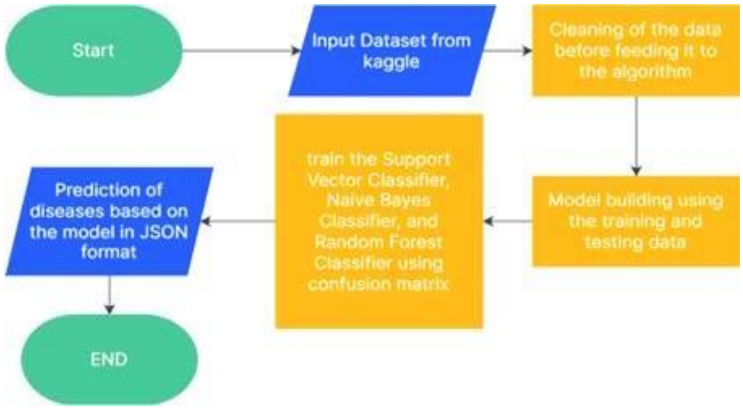
B. Nithya and V. Ilango, in [10] researched on "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 2017. The research paper "Predictive analytics in healthcare using machine learning tools and techniques" by Nithya and Ilango (2017) discusses the potential applications of machine learning techniques in healthcare. The authors argue that the use of machine learning in healthcare can lead to improved patient outcomes, reduced healthcare costs, and better resource allocation. The authors conclude that machine learning techniques have significant potential in healthcare and can lead to improved patient outcomes and reduced healthcare costs. The authors also note that the use of machine learning in healthcare requires collaboration between healthcare professionals and data scientists.

The paper presents a case study in which the authors used machine learning techniques to predict hospital read- missions. The authors used a dataset of patient records from a hospital and trained and tested machine learning models to predict the likelihood of hospital readmission. The authors found that their models were able to predict hospital readmissions with high accuracy.

The authors conclude that machine learning techniques have significant potential in healthcare and can lead to improved patient outcomes and reduced healthcare costs. The authors also note that the use of machine learning in healthcare requires collaboration between healthcare professionals and data scientists.

Overall, the paper provides a valuable contribution to the field of machine learning and healthcare by discussing the potential applications of machine learning techniques in healthcare. The authors' case study provides evidence for the effectiveness of machine learning in predicting health outcomes. However, the paper is limited by its focus on a single case study and the absence of external validation of the proposed approach. Future research should explore the effectiveness of machine learning techniques in other healthcare settings and populations.

IV. PROPOSED MODEL FLOW



V METHODOLOGY

Using machine learning techniques, the following steps are taken to create a disease prediction model:

- **Data Gathering:** The initial stage in creating a disease prediction model is to gather pertinent data. This could contain medical records, demographic data, genetic information, and data from medical imaging. For the machine learning algorithm to be trained effectively, it's critical to make sure the data is of high quality and contains a significant number of both positive and negative cases.

- **Preprocessing of Data**: After data has been gathered, it must be processed to ensure that it is in a format that is appropriate for machine learning algorithms. This could entail feature selection, normalization, and data cleansing. The goal is to prepare the data so that the machine learning algorithm may perform as well as possible.

- **Model Option:** The following stage is choosing the best machine learning algorithm for the problem at hand. There are several methods available, including decision trees, random forests, support vector machines, and neural networks. The choice of algorithm will depend on the type of data and the nature of the disease being predicted.

- **Model Training**: After choosing a machine learning algorithm, the model must be trained using the preprocessed data. This entails feeding the algorithm a subset of the data and modifying the algorithm's settings to improve performance. The model is trained repeatedly until a desirable level of accuracy is reached.

- **Model Evaluation:** After the model has been trained, it must be assessed using an additional set of data that was not utilized in training. This makes it more likely that the model will generalize to new data and not overfit to the training set of data. The performance of the model can be assessed using a variety of metrics, including accuracy, precision, recall, and F1-score.

Model optimization: Depending on the evaluation's findings, the model might need to be tweaked to perform better. This could entail changing the algorithm's inputs, including extra features, or gathering more information.

Model Deployment: Following model optimization, the model may be used in clinical settings. This entails incorporating the model into current healthcare systems and instructing medical personnel on its efficient application. In conclusion, developing a disease prediction model using machine learning techniques entails gathering and preprocessing data, choosing a suitable algorithm, train- ing and assessing the model, assessing its performance, and implementing the model in clinical practice.

VI WORKING

Dataset Preparation: To begin with, the dataset is loaded using the pandas library. The null column is dropped and a bar plot is used to check whether the target columnis balanced or not. The dataset is a clean one, with no null values and all features consisting of 0s and 1s. The target column, however, is of object datatype, so a label encoder is used to convert it into a numerical datatype. The label encoder assigns a unique index to each label and converts them into numerical form.

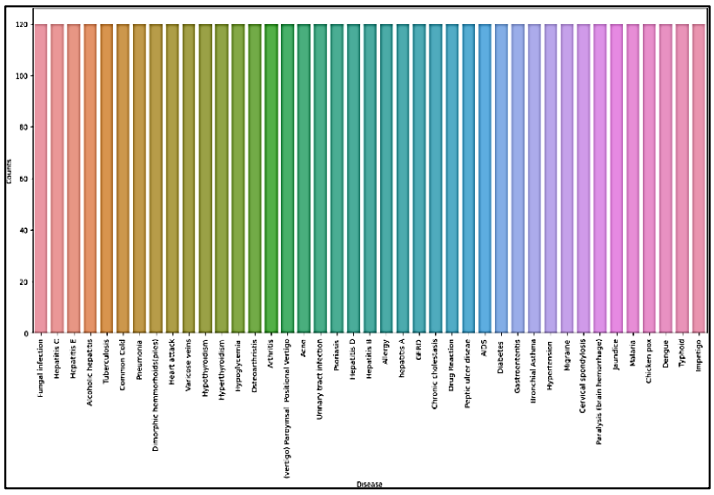
- **Train-Test Split:** After cleaning and encoding the data, it is split into 80:20 format for training and testing the model, respectively. The training data is used to trainthe models while the testing data is used to evaluate the performance of the models.

- **Model Building**: K-Fold cross-validation is used to evaluate the performance of the Support Vector Classifier, Gaussian Naive Bayes Classifier, and Random ForestClassifier. These models are briefly described, with Sup- port Vector Classifier being a discriminative classifier, Gaussian Naive Bayes Classifier being a probabilistic ma- chine learning algorithm, and Random Forest Classifier being an ensemble learning-based supervised machinelearning classification algorithm.

- **Using K-Fold Cross-Validation for Model Selection**: All three machine learning models perform well in cross-validation, and their mean scores are high. To build a robust model, the predictions of all three models are combined by taking their mode. This ensures that even if one of the models makes incorrect predictions, the other two models can make the correct ones, thereby making the final output more accurate.

- **Building Robust Classifier by Combining All Models:** After training the models on the entire train dataset, they are tested on the test dataset. The combined model accurately classifies all data points, indicating the effectiveness of the approach. Finally, a function is defined that takes symptoms separated by commas as input, uses the combined model to predict the disease based on the symptoms, and returns the predictions in a JSON format.

VII RESULTS AND DISCUSSION



From the above plot, we can observe that the dataset is a balanced dataset i.e. there are exactly 120 samples for each disease, and no further balancing is required. We can notice that our target column i.e. prognosis column is of object datatype, this format is not suitable to train a machine learning model. So, we will be using a label encoder to convert the prognosis column to the numerical datatype. Label Encoder converts the labels into numerical form by assigning a unique index to the labels. If the total number of labels is n, then the numbers assigned to each label will be between 0 to n-1.

Disease

```
In [3]: # Encoding the target value into numerical
        # value using LabelEncoder
        encoder = LabelEncoder()
        data["prognosis"] = encoder.fit_transform(data["prognosis"])

In [4]: X = data.iloc[:,:-1]
        y = data.iloc[:, -1]
        X_train, X_test, y_train, y_test =train_test_split(
        X, y, test_size = 0.2, random_state = 24)

        print(f"Train: {X_train.shape}, {y_train.shape}")
        print(f"Test: {X_test.shape}, {y_test.shape}")

        Train: (3936, 132), (3936,)
        Test: (984, 132), (984,)

In [5]: # Defining scoring metric for k-fold cross validation
        def cv_scoring(estimator, X, y):
            return accuracy_score(y, estimator.predict(X))

        # Initializing Models
        models = {
```

Now that we have cleaned our data by removing the Null values and converting the labels to numerical format, It's time to split the data to train and test the model. We will be splitting the data into 80:20 format i.e. 80% of the dataset will be used for training the model and 20% of the data will be used to evaluate the performance of the models.

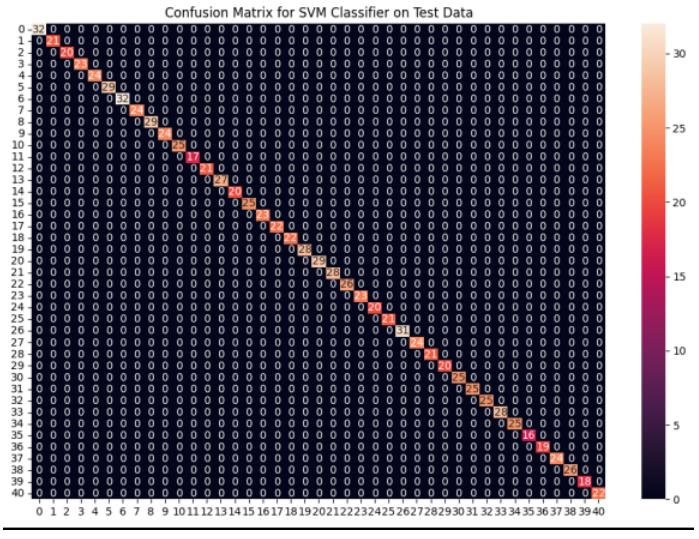
```
# Initializing Models
models = {
    "SVC":SVC(),
    "Gaussian NB":GaussianNB(),
    "Random Forest":RandomForestClassifier(random_state=18)
}

# Producing cross validation score for the models
for model_name in models:
    model = models[model_name]
    scores = cross_val_score(model, X, y, cv = 10,
                             n_jobs = -1,
                             scoring = cv_scoring)
    print("==="*30)
    print(model_name)
    print(f"Scores: {scores}")
    print(f"Mean Score: {np.mean(scores)}")
```

```
==========================================================
SVC
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
==========================================================
Gaussian NB
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
==========================================================
Random Forest
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
```

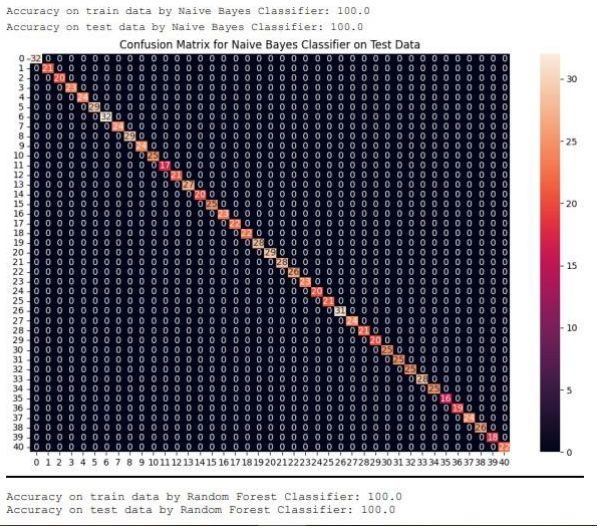
```
In [6]: # Training and testing SVM Classifier
```

```
Accuracy on train data by SVM Classifier: 100.0
Accuracy on test data by SVM Classifier: 100.0
```

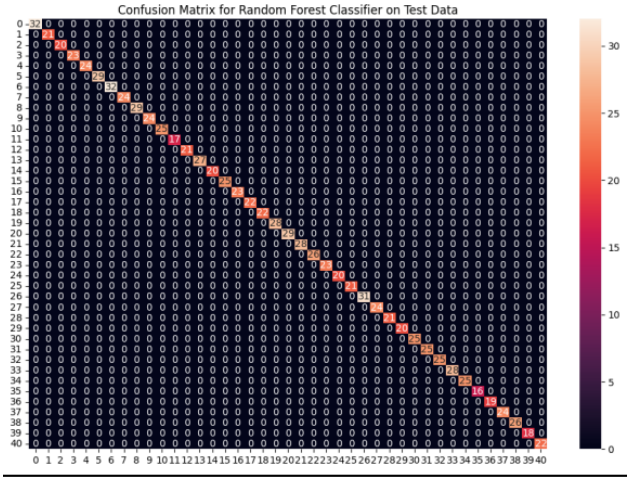
From the above output, we can notice that all our machine learning algorithms are performing very well and the mean scores after k fold cross-validation are also very high. To build a robust model we can combine i.e. take the mode of the predictions of all three models so that even one of the models makes wrong predictions and the other two make correct predictions then the final output would be the correct one. This approach will help us to keep the predictions much more accurate on completely unseen data. In the below code we will be training all the three models on the train data, checking the quality of our models using a confusion matrix, and then combine the predictions of all three models.



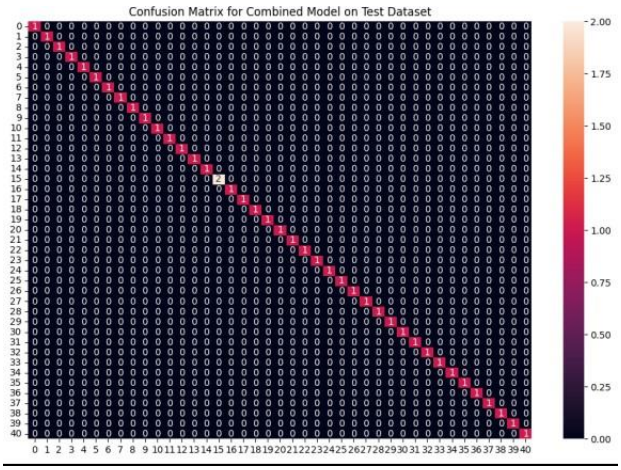
Accuracy on train data by Naive Bayes Classifier: 100.0
Accuracy on test data by Naive Bayes Classifier: 100.0



Accuracy on train data by Random Forest Classifier: 100.0
Accuracy on test data by Random Forest Classifier: 100.0



From the above confusion matrices, we can see that the models are performing very well on the unseen data. Now we will be training the models on the whole train data present in the dataset that we downloaded and then test our combined model on test data present in the dataset.



Accuracy on Test dataset by the combined model: 100.0

We can see that our combined model has classified all the data points accurately. We have come to the final part of this whole implementation, we will be creating a function that takes symptoms separated by commas as input and outputs the predicted disease using the combined model based on the input symptoms.

TEST CASE 1:

```
    predictions = {
        "rf_model_prediction": rf_prediction,
        "naive_bayes_prediction": nb_prediction,
        "svm_model_prediction": nb_prediction,
        "final_prediction":final_prediction
    }
    return predictions

# Testing the function
print(predictDisease("Itching,Skin Rash,Nodal Skin Eruptions"))
```

```
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature names
  warnings.warn(
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but GaussianNB was fitted with feature names
  warnings.warn(
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but SVC was fitted with feature names
  warnings.warn(
C:\Users\mysel\AppData\Local\Temp\ipykernel_10676\3148801054.py:37: FutureWarning: Unlike other reduction functions (e.g. `skew`, `kurtosis`), the default behavior of `mode` typically preserves the axis it acts along. In SciPy 1.11.0, this behavior will change: the default value of `keepdims` will become False, the `axis` over which the statistic is taken will be eliminated, and the value None will no longer be accepted. Set `keepdims` to True or False to avoid this warning.
  final_prediction = mode([rf_prediction, nb_prediction, svm_prediction])[0][0]
{'rf_model_prediction': 'Fungal infection', 'naive_bayes_prediction': 'Fungal infection', 'svm_model_prediction': 'Fungal infection', 'final_prediction': 'Fungal infection'}
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\scipy\stats\_stats_py.py:110: RuntimeWarning: The input array could not be properly checked for nan values. nan values will be ignored.
```

TEST CASE 2:

```
ata.org/docs/reference/api/pandas.DataFrame.mode.html.
  final_prediction = mode([rf_prediction, nb_prediction, svm_prediction])[0][0]
In [9]: print(predictDisease("Continuous Sneezing,Shivering,Chills,Joint Pain,Fatigue"))
{'rf_model_prediction': 'Allergy', 'naive_bayes_prediction': 'Allergy', 'svm_model_prediction': 'Allergy', 'final_prediction': 'Allergy'}
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature names
  warnings.warn(
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but GaussianNB was fitted with feature names
  warnings.warn(
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but SVC was fitted with feature names
  warnings.warn(
C:\Users\mysel\AppData\Local\Temp\ipykernel_10676\3148801054.py:37: FutureWarning: Unlike other reduction functions (e.g. `skew`, `kurtosis`), the default behavior of `mode` typically preserves the axis it acts along. In SciPy 1.11.0, this behavior will change: the default value of `keepdims` will become False, the `axis` over which the statistic is taken will be eliminated, and the value None will no longer be accepted. Set `keepdims` to True or False to avoid this warning.
  final_prediction = mode([rf_prediction, nb_prediction, svm_prediction])[0][0]
c:\users\mysel\appdata\local\programs\python\python38\lib\site-packages\scipy\stats\_stats_py.py:110: RuntimeWarning: The input array could not be properly checked for nan values. nan values will be ignored.
  warnings.warn("The input array could not be properly "
C:\Users\mysel\AppData\Local\Temp\ipykernel_10676\3148801054.py:37: DeprecationWarning: Support for non-numeric arrays has been deprecated as of SciPy 1.9.0 and will be removed in 1.11.0. `pandas.DataFrame.mode` can be used instead, see https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.mode.html.
  final_prediction = mode([rf_prediction, nb_prediction, svm_prediction])[0][0]
In [ ]:
```

As we can see, the program gives an accurate estimation of the disease a person may have if we give symptoms as input.

VIII CONCLUSION

In conclusion, disease prediction using machine learning is a promising area of research that has the potential to improve the accuracy and speed of diagnosing diseases. With the increasing availability of large datasets and advancements in machine learning algorithms, it is now possible to build robust and accurate models for disease prediction.In this approach, we have used K-Fold cross-validation to evaluate and select the best performing models for disease prediction. We have trained three different classifiers, namely,

Support Vector Classifier, Gaussian Naive Bayes Classifier, and Random Forest Classifier, and combined their predictions to build a more robust model. The final model has shown high accuracy in predicting diseases based on input symptoms. This approach can be further extended by incorporating more complex machine learning algorithms and larger datasets to improve the performance of the model. Overall, the use of machine learning in disease prediction has the potential to revolutionize healthcare by providing faster and more accurate diagnosis, leading to better treatment outcomes and ultimately improving patient care.

IX ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Dr. Aparna Mohanty, my faculty advisor, for her constant guidance and support throughout this research. Her expertise in the field of machine learning and her valuable insights have been instrumental in shaping this work.

I would also like to thank the Department of Electronics and Communication engineering at Vellore Institute of Technology for providing the necessary resources and infrastructure for carrying out this research.

Finally, I would like to extend my gratitude to all the participants who volunteered to share their medical data for this study. Without their contribution, this research would not have been possible.

X REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, ”Disease prediction by machine learning over big data from healthcare communities”, IEEE Access, vol. 5, no. 1, pp. 8869-8879, 2017.

[2] B. Qian, X. Wang, N. Cao, H. Li and Y.-G. Jiang, ”A relative similarity based method for interactive patient risk prediction”, Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070-1093, 2015.

[3] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang and C. Youn, ”Wearable 2.0: Enable human-cloud integration in next generation healthcare system”, IEEE Commun, vol. 55, no. 1, pp. 54-61, Jan. 2017.

[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan and A. Alamri, ”HealthCPS: Healthcare cyberphysical system assisted by cloud and big data”, IEEE Syst. J, vol. 11, no. 1, pp. 88-95, Mar. 2017.

[5] L. Qiu, K. Gai and M. Qiu, ”Optimal big data sharing approach for telehealth in cloud computing”, Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), pp. 184-189, Nov. 2016.

[6] Ajinkya Kunjir, Harshal Sawant and Nuzhat F. Shaikh, ”Data Mining and Visualization for prediction of Multiple Diseases in Healthcare”, IEEE big data analytics and computational intelligence, pp. 2325, Oct 2017.

[7] Shanthi Mendis, Pekka Puska and Bo Norrving, ”Global Atlas on

[8] S.U. Amin, K. Agarwal and R. Beg, ”Genetic neural network based data mining in prediction of heart disease using risk factors”, IEEE Conference on Information Communication Technologies (ICT), pp. 1227-31, 11-12 April 2013.

[9] S Palaniappan and R Awang, ”Intelligent heart disease prediction System using data mining techniques”, IEEE/ACS International Conference on Computer Systems and Applications AICCSA 2008, pp. 108115, March 31 2008-April 4 2008.

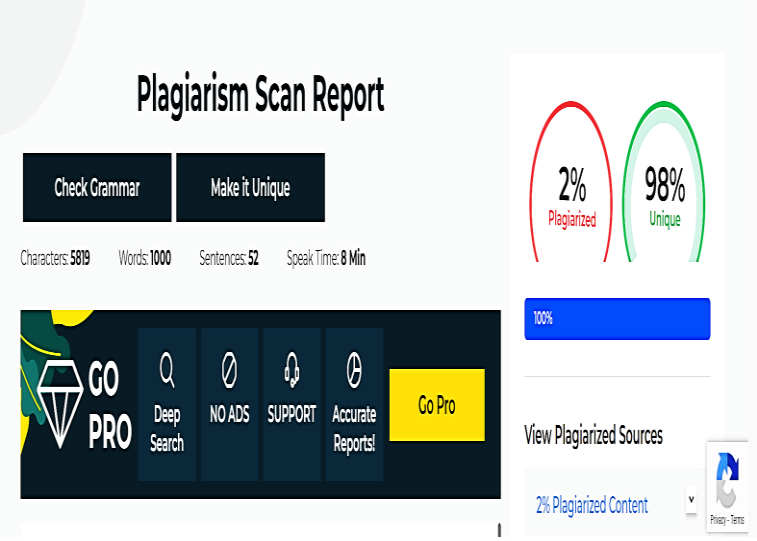
[10] B. Nithya and V. Ilango, ”Predictive Analytics in Health Care Using Machine Learning Tools and Techniques”, International Conference on Intelligent Computing and Control Systems, 2017.

[11] S. Leoni Sharmila, C. Dharuman and P. Venkatesan, ”Disease Classification Using Machine Learning Algorithms - A Comparative Study”, International Journal of Pure and Applied Mathematics, vol. 114, no. 6, pp. 1-10, 2017.

[12] Allen Daniel Sunny, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mo- han Ba and H Sarojadevi, ”Disease Diagnosis System By Exploring Machine Learning Algorithms”, International Journal of Innovations in Engineering and Technology (IJIET), vol. 10, no. 2, May 2018.

[13] Shraddha Subhash Shirsath, ”Disease Prediction Using Machine Learning over Big Data”, International Journal of Innovative Research in Science

XI PLAGARISM REPORT



CONFERENCE DETAILS

The above paper has been submitted in the

3rd ASIANCON 2023 ,Pune, India Technically Co-Sponsored by IEEE Bombay Section and Sponsored by AICTE New Delhi

https://asiancon.org/

https://cmt3.research.microsoft.com/ASIANCON2023/Submission/Index