Predicting disease from several symptoms using machine learning approach.

MD. Atikur Rahman[1], Tania Ahmed Nipa[2], Md. Assaduzzaman[3]

[1]*Student, Department of Computer Science and Engineering, Daffodil International University, DSC, Ashulia, Savar, Dhaka, Bangladesh*
[2]*Student, Department of Computer Science and Engineering, Daffodil International University, DSC, Ashulia, Savar, Dhaka, Bangladesh*
[3]*Lecturer, Department of Computer Science and Engineering, Daffodil International University, DSC, Ashulia, Savar, Dhaka, Bangladesh*

---***---

Abstract - *Humans are the most intelligent species on the earth and are very health conscious. The evolution of recent technologies like data science and machine learning has opened the trail for healthcare communities and medical establishments, to observe the diseases earliest as potential and it helps to supply higher patient care. For many medical organizations, disease prediction is very important for making the best possible healthcare decisions. Machine Learning is a field where we can develop a model to learn machines to make decisions on their own from real-time data and from past experience. We proposed a model to predict disease from some symptoms. So, in this experiment, we propose a new knowledge-based system for disease prediction using KNN, SVM, NB, DT, RF, and LR for data modeling and we got maximum 98.36 percent accuracy from the KNN algorithm. This paper is planned to develop multi-disease prediction using the machine learning concept. Our main contribution is to implement feature engineering and standard scal-ing to optimize our algorithm and better performance. The decision pro-posed support system for disease diagnosis might be implemented using the suggested methodology, aiding doctors in their work and enhancing patient outcomes. So, future studies will concentrate on increasing the dataset to cover a wider range of patient demographics and improving the machine learning algorithms to better prediction accuracy.*

Key Words: **Machine Learning, Disease, Syndromes**

1.INTRODUCTION

Machine learning is the programming field of computers where computer systems learn from data and experience. Nowadays, demands for health-related information are transforming information-seeking behavior, as demonstrated globally. Finding accurate health information online on symptoms, diagnoses, and treatments can just be time-consuming and expensive for many people. Today, billions of searches are performed daily, and sometimes the results are significant, and sometimes they are not. Such search terms generate thousands of results linked to medical advice. Diseases and health-related problems like Pneumonia, dengue, AIDS, Diabetes, Hepatitis, Jaundice, Arthritis, etc. Cause a major effect on someone's health and sometimes might also come to death if they ignore[1]. People usually want to know if their symptoms are indicative of any serious diseases. However, they cannot access any tools that would provide them with precise information. Mainly machine Learning technology gives us a superior platform in the medical field so that healthcare issues will be solved effectively[10]. From the Centers for Medicare and Medicaid Services, 50% of Americans have multiple chronic diseases with a total US healthcare expenditure in 2016 being about $3.3 trillion, which amounts to $10,348 per person in the US[11]. This project intends to give them the tools they need to tell end users about disease prediction. If review mining can be used to create a prediction system for physicians and medicine, a lot of time will be saved. Understanding complex medical terms, such as scientific names, might be challenging while using this type of system's user interface. The user is confused by the wealth of medical knowledge on numerous symptom categories that are provided. This system's objective is to change to meet the particular user interaction requirements of the health area. A crucial part of treatment is using symptoms to anticipate sickness. In our experiment, we give an effort to accurately predict a disease by examining the patient's symptoms.

In the field of health informatics, machine learning becomes more popular to diagnosis, prognosis, personalized medicine and identify any disease by using some machine leaning technique. So, if we can predict any disease at it's early stage, it will be more easy to give treatment to the patient by any medical servant. Overall, machine learning has the potential to change the medical industry by giving medical experts strong tools to more precisely and effectively predict, diagnose, and treat diseases.

2.LITERATURE REVIEW

Grampurohit, S. and Sagarnal, C., [1] Via assisting doctors in early clinical diagnosis and prediction, a classification method that was constructed employing machine learning techniques was intended to greatly aid inside this settlement for mental well-being challenges. One representative collection containing 49201 patient records having diagnosis covering 40 diseases were selected for the study. The multiple regression was made up of 42 disorders. There must have been 95 of 132 independent factors (symptomatology) that were clearly related to diseases. In just this research study, a sickness estimation system was originally developed using machine learning algorithms like Classifier, Random Forest classifier, and the Naive Bayes classifier was exhibited. In just this study, 3 systems' effectiveness on a health file is

thoroughly compared, with each algorithm producing results with just a reliability of upwards to 94%
Marouane Fethi Ferjani's [2] Through examining key metrics, these research seeks to find trends across different based On supervised model types towards sickness detection. Regarding classifiers, the K-Nearest Neighbor, Decision Tree (DT), and Naïve Bayes classifier (NB) algorithms attracted the most interest (KNN). Svm Classifier (SVM) is indeed the best at detecting parkinson's, as according to studies. Using Logistic Regression (LR) performed extremely well enough in terms of predicting heart conditions. Lastly, recommendations were performed employing Convolutional Neural Networks (CNN) and Random Forest (RF), accordingly, both precise specific diseases of infectious symptoms.

Anant Agrawal, Harshit Agrawal, Shivam Mittal, and Mradula Sharma declared [3] About this research, researchers propose a hybrid machine learning framework composed of support vector machines and evolutionary algorithms. Data sets for the liver, diabetes, and heart, that everything retrieved either from a reputed university were utilized to assess our algorithm. On the data sets, researchers evaluated their algorithms. These were 75.4% effective on the Cleveland Heart dataset. By maintaining all of the parameters, researchers were able to cut the feature count from Thirteen to Ten with sustaining overall compromise on accuracy as evaluated by the SVM alone across the entire data. For such liver datasets, they achieved the highest precision (78.6%). By minimizing the number of components, they were able to achieve marginally lower accurateness, however, they were both still well within acceptable parameters.

Chauhan, R.H., Naik, D.N., Halpati, R.A., Patel, S.J., and Prajapati, M.A., discussed [4] This test measures the patient's signs as input and determined the risk that the sickness may arise. Utilizing a decision-tree classifier, disease prediction is actually achieved. The frequency of an illness is computed by a decision tree classifier. Appropriate diagnosis data management improves in earlier disease classification and clinical management as big data demand increases inside the healthcare and biomedical industries.

Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., wrote [5] it created a methodology to predict diseases employing multiple Machine learning techniques. Upwards of 235 diseases were included in the dataset that has been examined. This assessment system includes an outcome as such sickness which a person might well be encountering in light of the symptoms, age, and gender of the patient. Through comparing the different algorithms, a balanced KNN algorithm generated the greatest outcomes. This balanced KNN algorithm had a predictive performance around 93.6%. If an illness is identified earlier, our diagnostics program may play the role of a doctor, enabling appropriate care as well as the possibility of ensuring safety.designations.

3.METHODOLOGY

So, In this part, we have discussed some important phase So, In this part, we have discussed some important phase that is given below :

- Dataset Description.
- Data Preprocessing and exploration.
- Feature Engineering.
- Model Selection.
- System Architecture.
- Model Evaluation.
- Result and Discussion.

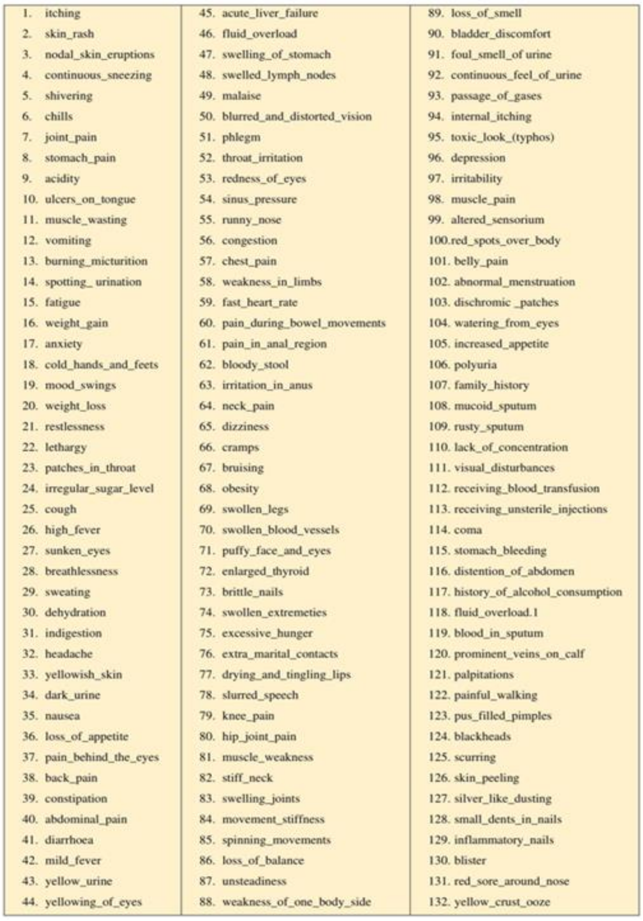
3.1 DATASET COLLECTION:

We collect the data from Kaggle[16], which has 132 syndromes from 4920 pa-tient data and we have taken into consideration consists of 132 symptoms and 41 disorders. We intend to create a model that incorporates user-provided symptoms and predicts the disease based on the 4920 patient data.

3.2 DATASET DESCRIPTION:

From the dataset, here is all syndromes that we analyzed for this experiment. This 132 syndrome is our independent variable. Depending on those syndromes we target the prognosis column to predict disease.

Table-1: The table shows the all syndromes from our dataset.



1. itching
2. skin_rash
3. nodal_skin_eruptions
4. continuous_sneezing
5. shivering
6. chills
7. joint_pain
8. stomach_pain
9. acidity
10. ulcers_on_tongue
11. muscle_wasting
12. vomiting
13. burning_micturition
14. spotting_ urination
15. fatigue
16. weight_gain
17. anxiety
18. cold_hands_and_feets
19. mood_swings
20. weight_loss
21. restlessness
22. lethargy
23. patches_in_throat
24. irregular_sugar_level
25. cough
26. high_fever
27. sunken_eyes
28. breathlessness
29. sweating
30. dehydration
31. indigestion
32. headache
33. yellowish_skin
34. dark_urine
35. nausea
36. loss_of_appetite
37. pain_behind_the_eyes
38. back_pain
39. constipation
40. abdominal_pain
41. diarrhoea
42. mild_fever
43. yellow_urine
44. yellowing_of_eyes
45. acute_liver_failure
46. fluid_overload
47. swelling_of_stomach
48. swelled_lymph_nodes
49. malaise
50. blurred_and_distorted_vision
51. phlegm
52. throat_irritation
53. redness_of_eyes
54. sinus_pressure
55. runny_nose
56. congestion
57. chest_pain
58. weakness_in_limbs
59. fast_heart_rate
60. pain_during_bowel_movements
61. pain_in_anal_region
62. bloody_stool
63. irritation_in_anus
64. neck_pain
65. dizziness
66. cramps
67. bruising
68. obesity
69. swollen_legs
70. swollen_blood_vessels
71. puffy_face_and_eyes
72. enlarged_thyroid
73. brittle_nails
74. swollen_extremeties
75. excessive_hunger
76. extra_marital_contacts
77. drying_and_tingling_lips
78. slurred_speech
79. knee_pain
80. hip_joint_pain
81. muscle_weakness
82. stiff_neck
83. swelling_joints
84. movement_stiffness
85. spinning_movements
86. loss_of_balance
87. unsteadiness
88. weakness_of_one_body_side
89. loss_of_smell
90. bladder_discomfort
91. foul_smell_of urine
92. continuous_feel_of_urine
93. passage_of_gases
94. internal_itching
95. toxic_look_(typhos)
96. depression
97. irritability
98. muscle_pain
99. altered_sensorium
100.red_spots_over_body
101. belly_pain
102. abnormal_menstruation
103. dischromic _patches
104. watering_from_eyes
105. increased_appetite
106. polyuria
107. family_history
108. mucoid_sputum
109. rusty_sputum
110. lack_of_concentration
111. visual_disturbances
112. receiving_blood_transfusion
113. receiving_unsterile_injections
114. coma
115. stomach_bleeding
116. distention_of_abdomen
117. history_of_alcohol_consumption
118. fluid_overload.1
119. blood_in_sputum
120. prominent_veins_on_calf
121. palpitations
122. painful_walking
123. pus_filled_pimples
124. blackheads
125. scurring
126. skin_peeling
127. silver_like_dusting
128. small_dents_in_nails
129. inflammatory_nails
130. blister
131. red_sore_around_nose
132. yellow_crust_ooze

Table-2: All the syndromes in prognosis.

1. Fungal infection
2. Allergy
3. GERD
4. Chronic cholestasis
5. Drug Reaction
6. Peptic ulcer disease
7. AIDS
8. Diabetes
9. Gastroenteritis
10. Bronchial Asthma
11. Hypertension
12. Migraine
13. Cervical spondylosis
14. Paralysis (brain hemorrhage)
15. Jaundice
16. Malaria
17. Chickenpox
18. Dengue
19. Typhoid
20. hepatitis A
21. Hepatitis B
22. Hepatitis C
23. Hepatitis D
24. Hepatitis E
25. Alcoholic hepatitis
26. Tuberculosis
27. Common Cold
28. Pneumonia
29. Dimorphic hemorrhoids(piles)
30. Heart attack
31. Varicose veins
32. Hypothyroidism
33. Hyperthyroidism
34. Hypoglycemia
35. Osteoarthritis
36. Arthritis
37. (vertigo) Paroxysmal Positional Vertigo
38. Acne
39. Urinary tract infection
40. Psoriasis
41. Impetigo

3.3 Data preprocessing and exploration:

Data pre-processing is an approach that changes the raw data or encodes the data into a form that the algorithm can quickly decode. The preprocessing methods applied in the study that is being presented in given below:

Data cleaning: Data cleaning is an important method for data pre-processing. Data is made clean through procedures including filling in missing values and removing null values which eliminate the data's discrepancies.

Data reduction: When working with large databases, analysis becomes challenging. That's why we eliminate some independent values that will not or have little effect on the target variable. So, 95 out of 132 symptoms that are closely related to the diseases are chosen for this research.
So for betterment, we have made several data visualizations to check the case of these factors.

In the dataset, we got several null values present in the columns. For modeling purposes, we decided to remove all the null values from each column and we converted all object values into categorical integer values.
From the given dataset, we were selected four target columns, which include the results from prognosis.
Now, here is the correlation matrix diagram from our experiment.

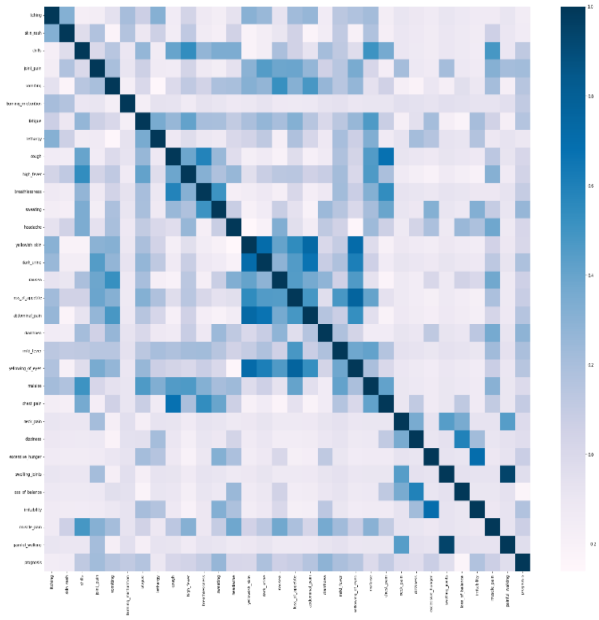


Chart-1 : Correlation Matrix.

So, the correlational matrix is a table that shows the correlation coefficients between two symptoms.

3.4 Feature Engineering:

After preprocessing the dataset, we got 132 symptoms from 4920 patients. Then we use some method to predict the accuracy but the was 100 percent which was overfitted. So, in this case, we use feature engineering to remove low-importance features. After removing the low-importance feature we got 33 symptoms from 4920 patients. For example :

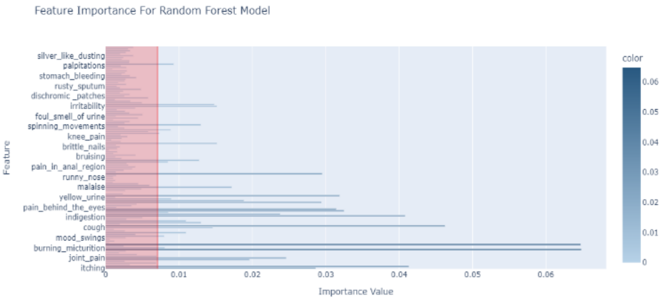


Chart-2: Feature Importance for Random Forest Model.

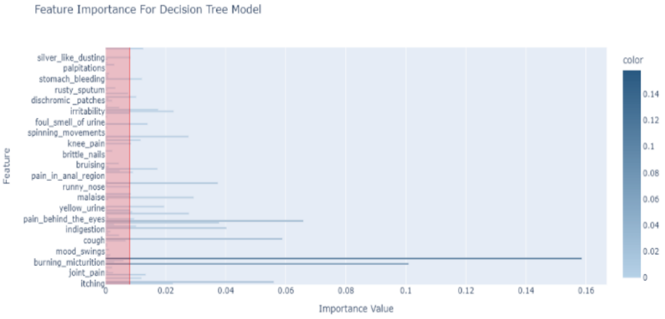


Chart-3: Feature Importance for Decision Tree Model.

Lastly, after using feature engineering, we get the final data shape and we split the dataset into train and test datasets.

3.5 Model Selection:

We use six algorithms to predict diseases from the trained model.

- **K-Nearest Neighbor (KNN):**
- **Support Vector Machine. (SVM)**
- **Gaussian Naive Bayes (GNB)**
- **Decision Tree Classifier. (DT)**
- **Random forest Classifier. (RF)**
- **Logistic Regression. (LR)**

3.6 Working Procedure:

We analyze some historical data depending on syndromes. Then we input all data after cleaning the dataset to learn every model to predict the outcome.

Here are some models that we use in this project:

K- NEAREST NEIGHBORS – k-NN is a supervised learning classifier, non-parametric, which uses proximity to predictions the group of an individual data point.
SUPPORT VECTOR MACHINE – SVM is a type of deep learning supervised algorithm for the classification or regression of data groups.

GAUSSIAN NAÏVE BAYES – GNB is a probabilistic classification algorithm based on applying Bayes' theorem with strong independence assumptions.

DECISION TREE – A decision tree is a particularly particular sort of probabilistic tree that allows users to choose an approach to take a decision. It is a supervised learning method that is used for both classification and regression applications.

RANDOM FOREST – Leo Breiman and Adele Cutler are the creators of the widely used machine learning technique known as random forest, which mixes the output of several decision trees to produce a single outcome.

LOGISTIC REGRESSION – A statistical analysis approach called logistic regression uses previous observations from a data set to predict a binary result, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable.

3.7 System Architecture:

In the given figure shows the working procedure of our system. In the system firstly we select the dataset then we preprocess the whole dataset, 2ndly we use feature engineering to remove all the low-importance features from the chosen dataset. 3rdly we clean the whole dataset after that we divide the dataset into train and test set. Lastly, we choose six classifiers to predict the disease.

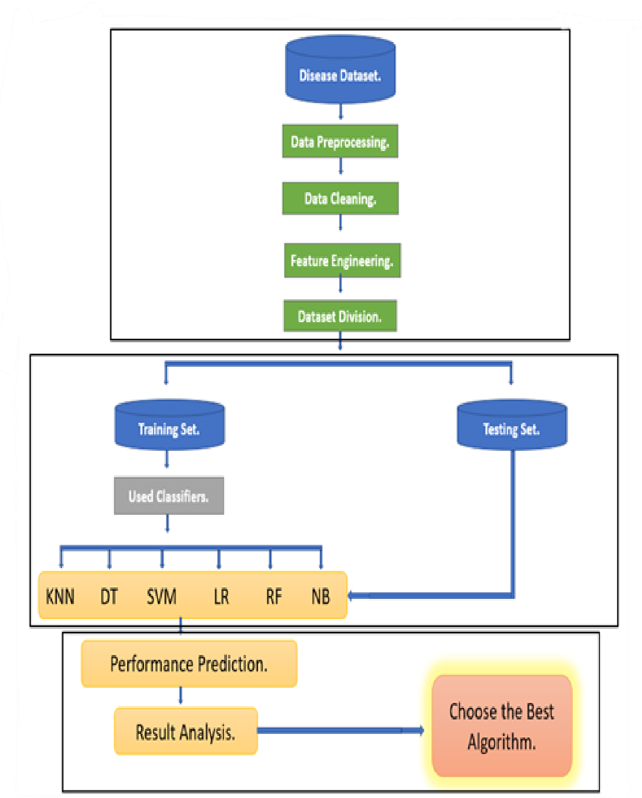


Chart-4 : System Architecture.

3.8 Evolution Method:

In this experiment, we calculate the performance of evolution. Firstly we represent TP, TN, FP, and FN, which represent True Positive, True Negative, False Positive, and False Negative respectively. Now, we will calculate the four measurements: recall, precision, accuracy, and F1 score as given below :

Accuracy: $$\frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: $$\frac{TruePositive}{TruePositive + FalsePositive} \quad (2)$$

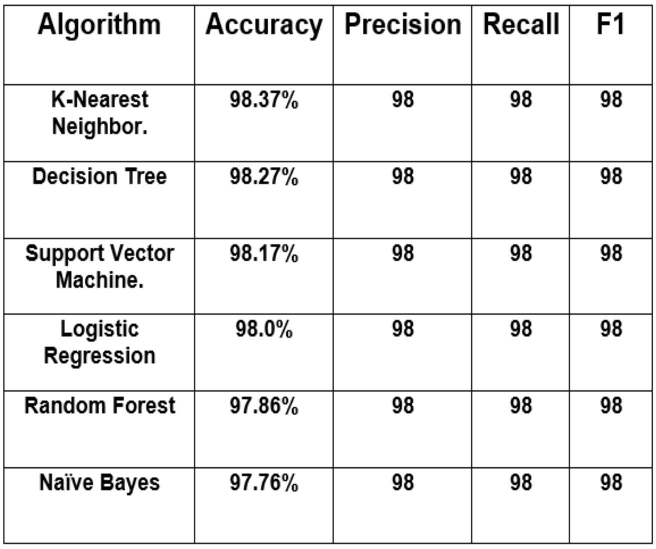
Recall: $$\frac{TruePositive}{TruePositive + FalseNegative} \quad (3)$$

F1-Score : $$\frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

4. RESULT AND DISCUSSION.

In this work we use six algorithms to predict the accuracy of the model and calculate the classification report of each algorithm. So, the table shows the result obtained from our experiment.

Table-3: The table shows the accuracy and classification report of each algorithm.



Algorithm	Accuracy	Precision	Recall	F1
K-Nearest Neighbor.	98.37%	98	98	98
Decision Tree	98.27%	98	98	98
Support Vector Machine.	98.17%	98	98	98
Logistic Regression	98.0%	98	98	98
Random Forest	97.86%	98	98	98
Naïve Bayes	97.76%	98	98	98

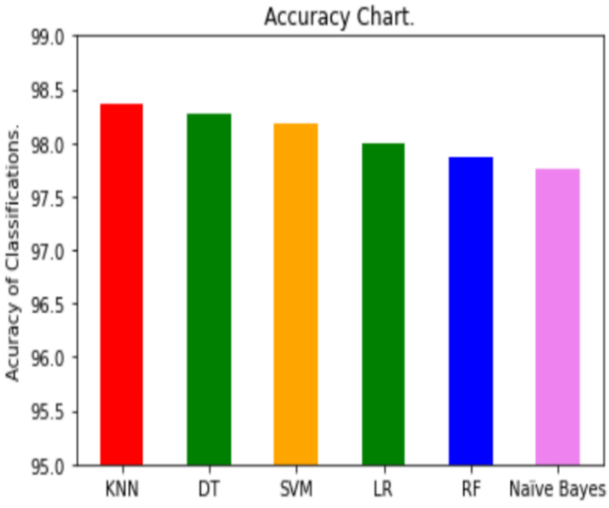


Chart -5: The chart shows the accuracy of six classifiers.

Here we use six machine learning models to predict the disease. Out of six models, we got 97% or more accuracy for all models. As shown in Figure-5 out of all the models, we got the highest accuracy for the KNN (K-Nearest Neighbor) that is 98.37% and the lowest accuracy for NB (Naive Bayes) which was 97.76%. Moreover, other models are DT, SVM, LR and RF Classifiers also the accuracy is accordingly 98.27%, 98.17%, 98.0%, and 97.86%. Lastly, Precision, recall, and F1 scores are 98% for all the models.

5. CONCLUSION.

This project aims to predict any disease with the given symptoms. So, the experiment that we organized in a way that system takes input which is symptoms from the user and generates output by predicting the actual disease. To sum up, the variety of aspects of the hospital data affects how accurately risk is predicted for diseases using risk modeling and we got maximum accuracy up to 98.37 percent from KNN algorithm.

REFERENCES

[1] Grampurohit, S. and Sagarnal, C., 2020, June. Disease prediction using machine learning algorithms. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.

[2] Ferjani, M.F., 2020. Disease Prediction Using Machine Learning. Bournemouth, England: Bournemouth University.

[3] Agrawal, A., Agrawal, H., Mittal, S. and Sharma, M., 2018, April. Disease prediction using machine learning. In Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) (pp. 26-27).

[4] Chauhan, R.H., Naik, D.N., Halpati, R.A., Patel, S.J. and Prajapati, M.A., 2008. Disease Prediction using Machine Learning. clinical reports, pp.783-7.

[5] Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., 2020. Disease prediction from various symptoms using machine learning. Available at SSRN 3661426.

[6] Dahiwade, D., Patle, G. and Meshram, E., 2019, March. Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

[7] Qian, B., Wang, X., Cao, N., Li, H. and Jiang, Y.G., 2015. A relative similarity based method for interactive patient risk prediction. Data Mining and Knowledge Discovery, 29(4), pp.1070-1093.

[8] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), pp.1-16.

[9] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, pp.8869-8879.

[10] Pingale, K., Surwase, S., Kulkarni, V., Sarage, S. and Karve, A., 2019. Disease prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 6, pp.831-833.

[11] Kohli, P.S. and Arora, S., 2018, December. Application of machine learning in disease prediction. In 2018 4th International conference on computing communication and automation (ICCCA) (pp. 1-4). IEEE.

[12] Liu, J., Zhang, Z. and Razavian, N., 2018, November. Deep ehr: Chronic disease prediction using medical notes. In Machine Learning for Healthcare Conference (pp. 440-464). PMLR.

[13] Jadhav, S., Kasar, R., Lade, N., Patil, M. and Kolte, S., 2019. Disease prediction by machine learning from healthcare communities. International Journal of Scientific Research in Science and Technology, pp.29-35.

[14] Vinitha, S., Sweetlin, S., Vinusha, H. and Sajini, S., 2018. Disease prediction using machine learning over big data. Computer Science & Engineering: An International Journal (CSEIJ), 8(1), pp.1-8.

[15] Vijiyarani, S. and Sudha, S., 2013. Disease prediction in data mining technique–a survey. International Journal of Computer Applications & Information Technology, 2(1), pp.17-21.

[16] https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning

[17] https://iq.opengenus.org/gaussian-naive-bayes/

[18] https://www.ibm.com/topics/random-forest

BIOGRAPHIES

Md. Atikur Rahman currently a undergraduate student till 2023 and studying in department of Computer Science and Engineering in Daffodil International University, Bangladesh. He works in the field of Machine Learning, Deep Learning, NLP & Artificial Intelligence also area of interest is health.

Tania Ahmed Nipa currently a student in the Department of Computer Science and Engineering at Daffodil International University. She have a genuine passion for web development and specifically specialize in the MERN stack. In addition to her technical skills, also she have a strong passion in the field of research.

Md. Assaduzzaman is a ex student of Daffodil International University. He is currently working as a Senior Lecturer at Daffodil International University. His interest in research area is Machine learning and NLP.