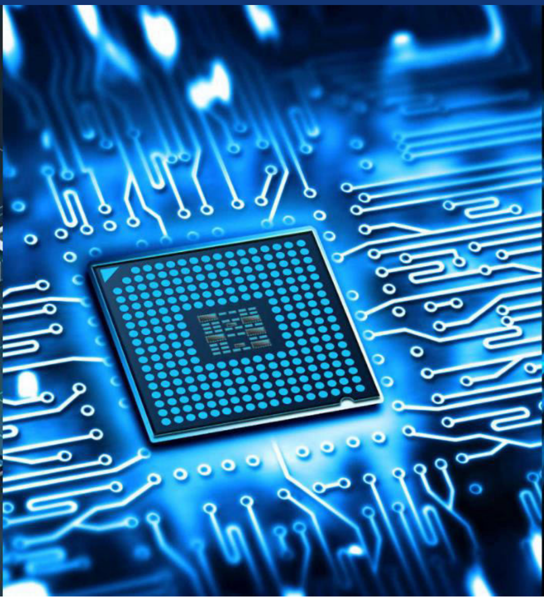
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.625

Volume 13, Issue 1, January 2025

www.ijircce.com ✉ ijircce@gmail.com ☎ +91-9940572462 ✆ +91 63819 07438

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

MEDIMATCH: Building and Deploying a Medicine Suggestion Application by Symptom-Driven Analysis

Mynamptati Sri Ranganadha Avinash, Sanjana S Acharya, Dabbara Asritha, Himanshu Sekhar Rout

B. Tech Student, Department of CSE, Presidency University, Bengaluru, India

B. Tech Student, Department of CSE, Presidency University, Bengaluru, India

B. Tech Student, Department of CSE, Presidency University, Bengaluru, India

Assistant Professor, School of CSE & IS, Presidency University, Bengaluru, India

ABSTRACT: A full drug recommendation system entitled MEDIMATCH that predicts a disease using refined algorithms of artificial intelligence based upon symptoms and assists the user about nutrition, exercising, drugs as well as about the safe ty practices. Using this, it approaches the error value at 94% level, with the best prediction of Medicine Recommendation System Datasets from the Kaggle datasets Other than being able to provide a forecasting result, the usage of visualization, like bipartite graphs and cosine similarity, increases the interpretability of and enhances users' confidence about disease and symptoms. The dataset was kindly provided with information shared via Noor YouTube channel and a GitHub repository that allows ethically and flexible usage, all being under an Apache 2.0 license. This collaborative effort shows the revolutionary potential of machine learning in customized medicine and public health solutions.

I. INTRODUCTION

With current developments in machine learning, health services can innovate within diagnosis and even treatment plans with innovative ideas on healthcare support through accurate, easily usable recommendation systems for physicians as they look towards improving their professional work in assisting optimum patient outcome together with accessible service. In recent years, these diagnostic systems highly rely on hand analysis which becomes both time and human-error-oriented. This can now be achieved using machine learning wherein an automated system can predict a disease based on symptoms and generate a complete care plan, along with dietary prescriptions, medications, and preventive measures.

Our proposed system, MEDIMATCH, aims to overcome the above challenges using machine learning to predict diseases accurately and provide recommendations holistically. The system is built using Medicine Recommendation System Dataset from Kaggle, which is a rich and diverse dataset meant for healthcare applications. By merging predictive analytics with actionable insights, MEDIMATCH aims to fill the gap between complex medical data and practical decision-making.

The current work is led by the author, Mr. Mynampati Sri Ranganadha Avinash, with Sanjana S Acharya and Dabbara Asritha as co-authors who helped in conducting the project. They were guided by Mr. Himanshu Sekhar Rout, Assistant Professor, School of CSE & IS, Presidency University, and finds his inspiration in the work by Noor Saeed, whose data and teaching resources form the bedrock of the dataset used here; it's, therefore, an example of a collaborative effort.

II. DATASET DESCRIPTION

Medicine Recommendation System Dataset is a structured dataset that is comprehensive for developing healthcare applications. It contains the following:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Symptom-severity.csv: It contains 132 symptoms with severity levels ranging between 1, which is the lowest level, and 7, which is the highest level. It is a granular level that can distinguish between very mild and severe presentations of diseases.
Training.csv: It is the biggest file in the dataset, having 4920 records and 133 columns that map symptoms into their relevant diseases. This is the primary source of data to train machine learning models.
Description.csv: It contains two columns, namely Disease and Description. The column describes in detail the descriptions of 41 diseases. Such descriptions are very important to educate the end-user and enhance the explainability of the system.
Diets.csv: Relates 41 diseases to their corresponding dietary recommendations, so it is quite useful for practical management of nutrition.
Medications.csv: The Recommended medications for all 41 diseases, so it is quite helpful in making a treatment plan.
Precautions_df.csv: Up to four precautionary measures for every disease, which will help reduce the risk of that disease and provide preventive care.
Symptoms_df.csv: Disease and symptom combinations map; it helps in knowing the disease pattern at a much more granular level.
Workout_df.csv: Workout routine designed for every disease, keeping physical well-being alongside medical treatment.
The Apache 2.0 license that governs the dataset ensures it is used ethically and adaptively, and appropriate credit is given to Noor Saeed for curating this valuable resource. His work through the AI with Noor YouTube channel and the GitHub repository significantly impacted the success of this project.

III. METHODOLOGY

3.1 Data Preprocessing

This step ensured checking for missing values and inconsistencies within the dataset; symptom-severity scores were normalized across features. "Disease" column underwent Label Encoding to make the disease names into numerical labels compatible with machine learning algorithms. The dataset was then split into the training set (70%) and testing set (30%) to assess how well the model generalizes. In addition, techniques of standardization were applied on the model in order to perform better in a manner that values of features were uniformly scaled.

3.2 Machine Learning Models

3.2.1 Support Vector Machine (SVM)

SVM is selected based on robustness for multi-class classification. The purpose of the algorithm was to find out the best distinguishing hyperplane between the classes in the high dimensional space. It utilized the kernel trick to handle the non-linear separability; hence, it mapped data into higher dimensional spaces. To optimize the performance of the model based on its hyperparameters, cross-validation was applied. The strength of the outputs from the SVM model was also very consistent at various runs and the accuracy level was strong.

3.2.2 CatBoost

CatBoost is used as a gradient boosting algorithm because it handles categorical data with very minimal preprocessing. It iteratively minimizes the loss function that gives 94% prediction accuracy and is better than SVM in terms of precision, recall, and F1 score. The categorical features are natively handled without overhead computation in training.

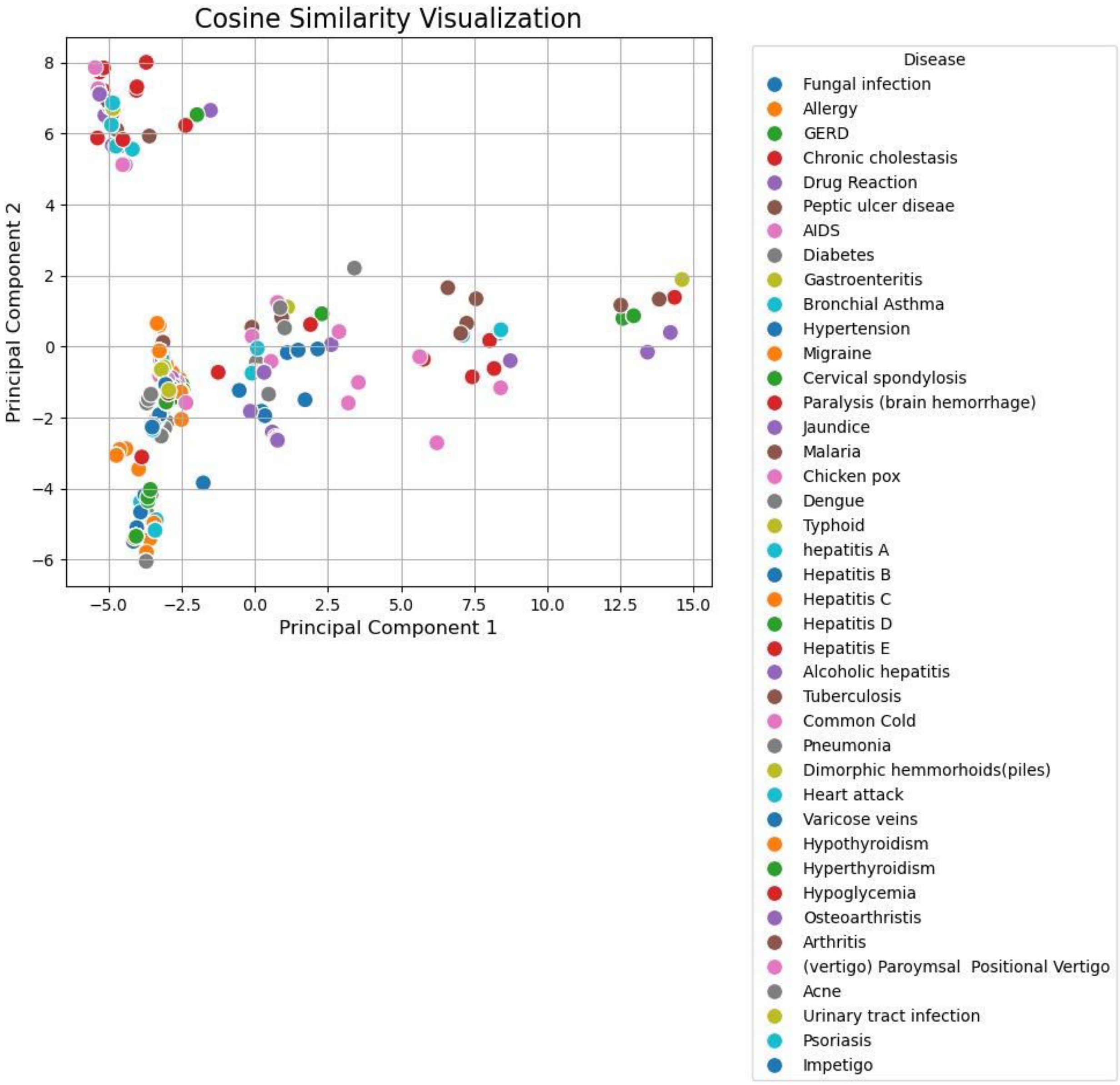
3.3 The Disease Prediction and Recommendations

The trained models were applied to the prediction of diseases corresponding to user-defined symptoms. It then fetched relevant details from auxiliary data files as given below:
Description.csv: Delineated explanatory details of the disease it had predicted.
Medications.csv: Provided lists with suggested treatments- this helped them know what types of medications were recommended for the patients.
Diets.csv: Referred to the Nutritional plans adjusted according to the treatment they suggested.
Precautions_df.csv: Generated actionable preventive measures concerning the risk mitigations.
Workout_df.csv: Many exercises for recovery and fitness
This hybrid recommendation system allows the users to obtain action and applicable advice about their condition.

IV. VISUALIZATION AND ANALYSIS

4.1 Cosine Similarity Matrix

To analyze the disease relationships in the context of symptom vectors, a cosine similarity matrix was constructed. The matrix was applied to quantify the similarity between diseases based on the symptoms that were embedded into high-dimensional vectors. Dimensions were reduced to two PC1 and PC2 using PCA. Scatter plot of PC1 vs. PC2 showed clusters of diseases, thereby pointing out the overlapping symptom profile groups. It helps find the similar symptomatically conditions which will increase the diagnostic accuracy and treatment planning.



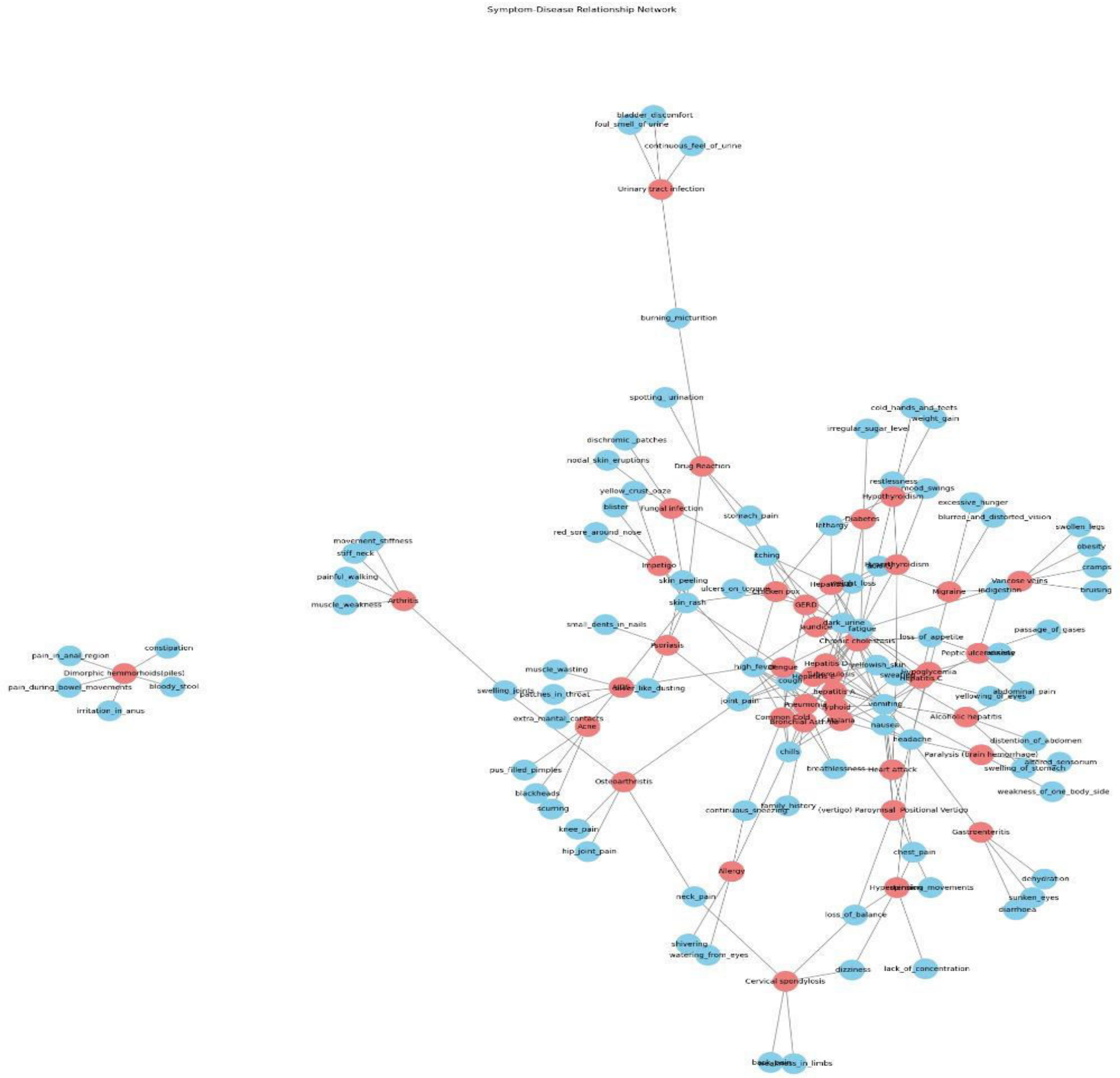
4.2 Visualisation of a Bipartite Graph

Utilising the NetworkX library, create a bipartite graph, thus graphically portraying the associations among symptoms and diseases. Here the nodes symbolise symptoms and diseases, respectively and edges depict an association between these symptoms and diseases. This form of graphing facilitates a much clearer depiction of the model.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

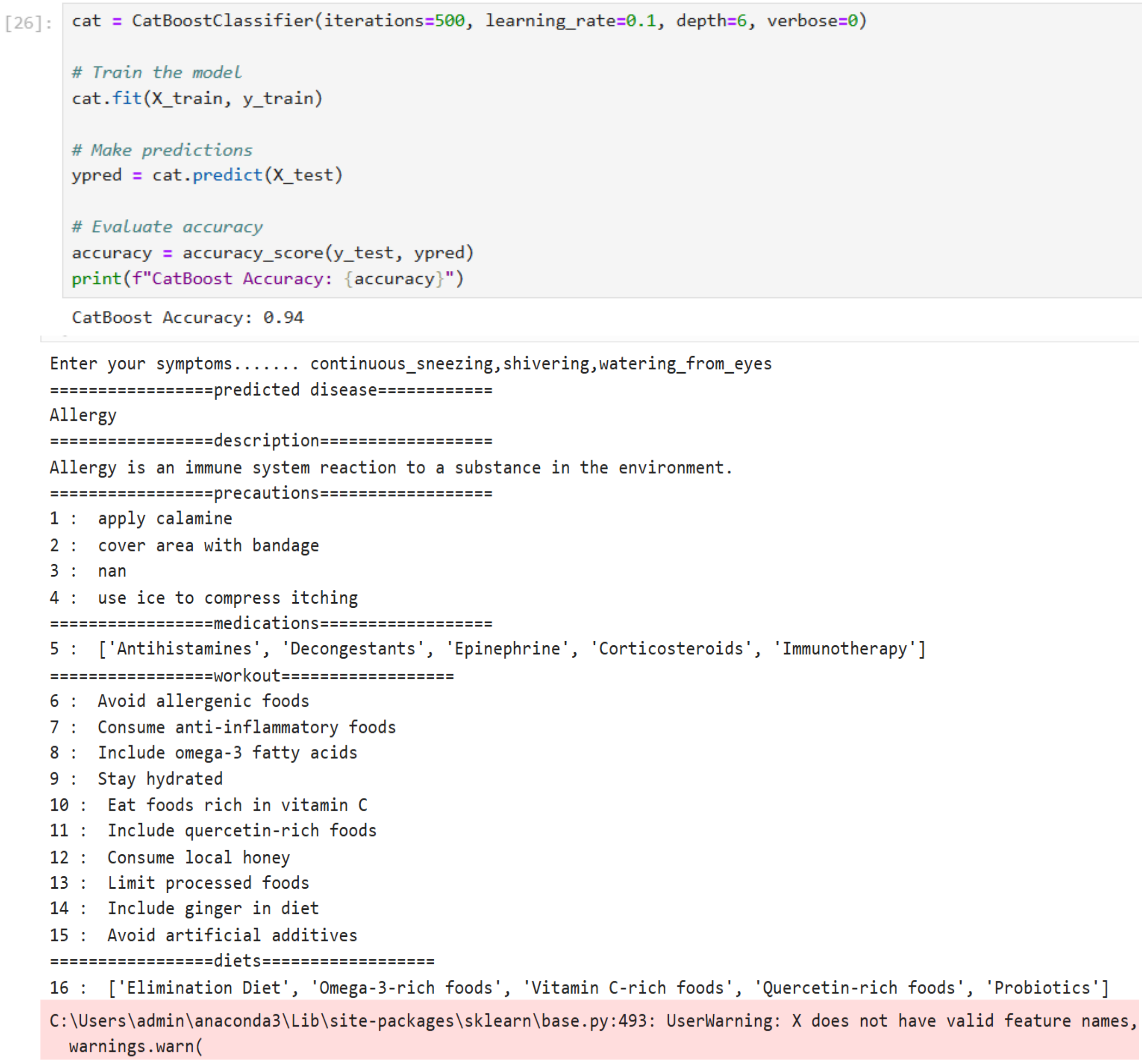
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



V. RESULTS

Model Accuracy. The SVM models produced accuracy results of 91% and the accuracy of CatBoost models was found to be 94%. Cosine Similarity Analysis: Scatterplots of the clusters of diseases were appropriately used to identify symptomatically similar conditions. Bipartite Graphs: It is here that with proper visualization regarding symptom-disease relationships, confidence of the users in the predictions made by the system improved.



```
[26]: cat = CatBoostClassifier(iterations=500, learning_rate=0.1, depth=6, verbose=0)

      # Train the model
      cat.fit(X_train, y_train)

      # Make predictions
      ypred = cat.predict(X_test)

      # Evaluate accuracy
      accuracy = accuracy_score(y_test, ypred)
      print(f"CatBoost Accuracy: {accuracy}")

      CatBoost Accuracy: 0.94
```

```
Enter your symptoms....... continuous_sneezing,shivering,watering_from_eyes
=================predicted disease============
Allergy
=================description=================
Allergy is an immune system reaction to a substance in the environment.
=================precautions==================
1 :  apply calamine
2 :  cover area with bandage
3 :  nan
4 :  use ice to compress itching
=================medications==================
5 :  ['Antihistamines', 'Decongestants', 'Epinephrine', 'Corticosteroids', 'Immunotherapy']
=================workout=================
6 :  Avoid allergenic foods
7 :  Consume anti-inflammatory foods
8 :  Include omega-3 fatty acids
9 :  Stay hydrated
10 :  Eat foods rich in vitamin C
11 :  Include quercetin-rich foods
12 :  Consume local honey
13 :  Limit processed foods
14 :  Include ginger in diet
15 :  Avoid artificial additives
=================diets==================
16 :  ['Elimination Diet', 'Omega-3-rich foods', 'Vitamin C-rich foods', 'Quercetin-rich foods', 'Probiotics']
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\base.py:493: UserWarning: X does not have valid feature names,
  warnings.warn(
```

VI. CONCLUSION

MEDIMATCH does pretty well in adequately demonstrating the medicine recommendation system capable of predicting diseases with high accuracy as well as providing actionable health recommendations. It does this while ensuring the system is transparent to the users so that trust is built. It depicted the fact that this scope of machine learning applied to the applications of health care can provide pertinent, interpretable, and informative solutions. Future applications of MEDIMATCH would be implemented even in real-time data, personal recommendations for the patient, and the broad scope of diseases amplifying the influence of this system into the field of personalized medicine.

VII. FUTURE WORK

Future Improvements of MEDIMATCH
The following are the future improvements that should be associated with the MEDIMATCH:• Patient-Specific Data Integration-Comprises demographic data, genetic background, and other lifestyle factors.

• Real-time Updates-Making the mechanism available for new health knowledge updates into the system. Addition of the suitable dataset is updated.
• Deployment-Distribution as an accessible web application or mobile interface.
• Explanable AI (XAI)-Interpretable Model Inference Using state-of-the art Explainability methods.

ACKNOWLEDGMENTS

The dataset for this project comes from Kaggle and is licensed under Apache 2.0. The authors would like to thank Noor Saeed on contributions, which include this data-set and guidance on offer through the AI with Noor YouTube Channel (https://youtu.be/1xHU20MgvqI?si=IdjQ9BIMp9-ZFrIx) and GitHub Repo (https://github.com/611noorsaeed). His efforts have significantly contributed to the success of this project.

REFERENCES

1. Kaggle: Medicine Recommendation System Dataset - https://www.kaggle.com/datasets/noorsaeed/medicine-recommendation-system-dataset
2. Noor Saeed, AI with Noor - https://youtu.be/1xHU20MgvqI?si=IdjQ9BIMp9-ZFrIx
3. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory.
4. Prokhorenkova, L., et al. (2018). CatBoost: unbiased boosting with categorical features.
5. Newman, M. (2010). Networks: An Introduction.
6. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python.
7. Cortes, C., & Vapnik, V. (1995). Support-vector networks.
8. Bishop, C. M. (2006). Pattern Recognition and Machine Learning.
9. He, K., et al. (2015). Deep Residual Learning for Image Recognition.
10. Rumelhart, D. E., et al. (1986). Learning representations by back-propagating errors.
11. Goodfellow, I., et al. (2016). Deep Learning.
12. Hastie, T., et al. (2009). The Elements of Statistical Learning.
13. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective.
14. Chollet, F. (2018). Deep Learning with Python.
15. Abadi, M., et al. (2016). TensorFlow: A system for large-scale machine learning.
16. Koller, D., & Friedman, N. (2009). Probabilistic Graphical Models.
17. Jiawei, H., et al. (2011). Data Mining: Concepts and Techniques.
18. LeCun, Y., et al. (1998). Gradient-based learning applied to document recognition.
19. Aggarwal, C. C. (2018). Neural Networks and Deep Learning.
20. Mitchell, T. M. (1997). Machine Learning.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com