International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Integrated Feature Selection and Hyperparameter Optimization for Multi-Label Classification of Medical Conditions

Suraj Kumar[1], Kukku Youseff[2]

[1]Lead Data Scientist
Email: *surajatiitb[at]gmail.com*

[2]Data Science Post Graduate
Email: *kukkuyouseff[at]gmail.com*

Abstract: ***The continuous evolution of data technologies in biomedical and healthcare domains has propelled the importance of precise medical data analysis for early disease recognition, improved patient care, and community services. However, the accuracy of such analyses encounters challenges in the face of incomplete medical data and regional variations in disease presentations, which can impact the precision of disease outbreak predictions. This paper addresses these challenges by investigating advanced techniques for feature selection and hyperparameter tuning to elevate the performance of a machine learning based Classifier in predicting medical conditions based on symptoms. Our methodology incorporates Recursive Feature Elimination (RFE), Mutual Information, Information Gain, and LASSO for comprehensive feature selection. The proposed approach aims make the machine learning classification more deployable in real world because the data set used has 132 symptoms for classifying 42 diseases, in real world we cannot ask for these many symptoms to a person that's where feature engineering and hyperparameter optimization is used. Though classification of the disease has achieved exceptionally high accuracy which is a clear case of overfitting, we incorporate feature engineering for more practical machine learning model. Thus, we can address the problem of overfitting and deliver a reliable diagnosis model by this approach.***

Keywords: Machine learning, Feature selection, Classification.

1. Introduction

The domain of computer-aided diagnosis (CAD) in the medical industry is continuously evolving. In the past, people seeking medical consultation would navigate the process of booking appointments, enduring wait times, and filling forms before finally meeting with a physician to discuss their symptoms. This traditional approach often induced anxiety throughout the waiting period, heightening concerns about one's health. However, in the contemporary technological era, individuals frequently resort to online searches for symptom analysis, intensifying their apprehension. The accuracy of a study is closely related to the quality of a data [2].

Contrary to the initial intent of technology to streamline healthcare processes and provide reassurance, it has, regrettably, become a prominent source of anxiety for many. This research endeavors to enhance computer-aided diagnosis to empower individuals with the ability to comprehend their ailments more effectively. The primary objective is to implement machine learning algorithms for symptom-based disease classification, thereby facilitating a more efficient and accurate diagnostic process for the general populace. This undertaking seeks to alleviate the burden on individuals by providing a reliable and accessible means of understanding their health conditions through advanced technological applications.

This work utilizes a data typically consists of 42 diseases which is the target variable and their symptoms are the independent variable. We have around 132 symptoms the data set is binary encoded meaning 1 means symptoms is present and 0 means symptoms

Objective of this work

- To investigate advanced techniques for feature selection and hyperparameter tuning.
- To enhance the performance of a machine learning classifier for multi-label classification of medical conditions.
- To utilize symptom-based data for disease prediction.
- To develop a more accurate and reliable diagnostic model.
- Assist in early disease recognition and improve patient care.

Machine Learning Model

Here we are dealing what is known as supervised learning we are given a labelled dataset where there are dependent variable and independent variables. We are going to classify the independent variable into multiple classes of disease. So, this model will help us to diagnosis the underlying disease based on symptoms to an extent. This model can be an inspiration for online consultation.

Machine learning techniques have emerged as invaluable tools in disease diagnosis, offering the potential to analyse large volumes of medical data efficiently and accurately. In the context of symptom-based disease classification, several machine learning algorithms have been widely applied:

- Logistic Regression: A classic algorithm used for binary classification, logistic regression models the probability of a certain class or event occurring.
- Random Forest: A powerful ensemble learning method, random forest constructs a multitude of decision trees during training and outputs the mode of the classes for classification.
- Decision Tree: Decision trees partition the data into subsets based on features at each node, making them

particularly interpretable for medical diagnosis tasks.
- Support Vector Machine (SVM): SVMs aim to find the hyperplane that best separates classes in the feature space, making them effective for both linear and nonlinear classification tasks.
- K-Nearest Neighbors (KNN): KNN relies on the proximity of data points in feature space, assigning a class label based on the majority class among its k nearest neighbors.

Feature engineering plays a crucial role in improving the performance of machine learning models, especially in medical diagnosis where the input data can be high-dimensional and noisy. By selecting the most relevant features and eliminating irrelevant or redundant ones, feature engineering helps enhance model interpretability, reduce overfitting, and improve prediction accuracy.

Hyperparameters are parameters whose values are set prior to the training process. Optimizing these hyperparameters can significantly impact the performance of machine learning models. Techniques such as grid search, random search, and Bayesian optimization are commonly employed to tune hyperparameters and improve model generalization.

2. Literature Review

There is significant amount of research done on disease diagnosis using machine learning and also there are works done on same data set [1].

"Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach" Talasila Bhanute.et al explores the integration of machine learning (ML) techniques in healthcare, specifically focusing on disease prediction through data mining methodologies. Utilizing a dataset of 4920 patient records with 41 diseases and 132 symptoms, the study evaluates the performance of three ML algorithms: Decision Tree Classifier, Light GBM, and Random Forest Classifier. Results indicate high accuracy rates, with the Random Forest Classifier achieving the highest score of 98.315%. Methodological insights into the algorithms' implementations, including Gradient-based One-Side Sampling and ensemble techniques, are provided. The review concludes by emphasizing the significant role of ML in healthcare data analysis, highlighting its potential for improving disease diagnosis and patient care [1]

M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang introduced a novel approach called the Convolutional Neural Network-based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm, designed to leverage both structured and unstructured data from hospital records. Notably, this study represents the first attempt to incorporate both data types within the realm of medical big data analytics. Comparative analysis with existing prediction methods demonstrates that our proposed algorithm achieves a prediction accuracy of 94.8%, outperforming conventional algorithms. Furthermore, the CNN-MDRP algorithm exhibits a faster convergence rate compared to the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm. [2]

M. Chen, Y. Hao, K. Hwang, L. Wang and L examined various data mining and machine learning techniques applied in healthcare, particularly focusing on disease prediction models based on symptoms. Studies explored methods such as Decision Trees, Random Forests, and Support Vector Machines to develop accurate predictive models. Feature selection techniques like Recursive Feature Elimination (RFE), Mutual Information, Information Gain, and LASSO were investigated to optimize model performance. Additionally, the review highlighted the importance of precise medical data analysis in early disease recognition and patient care. The integration of machine learning algorithms into healthcare systems aims to assist in predicting and diagnosing diseases at early stages, thereby improving patient outcomes [3].

Current research predominantly centres around constructing and deploying diverse models to enhance model accuracy. The primary emphasis lies on exploring various machine learning and deep learning architectures. However, there exists a gap in the literature concerning investigations into feature importance and the development of integrated models. While considerable effort has been devoted to refining the structure and performance of individual models, there is limited exploration into the significance of specific features and the synergistic integration of multiple models. Addressing this gap could yield insights into the critical factors influencing model predictions and facilitate the creation of more comprehensive and effective predictive frameworks.

3. Proposed Methodology

1) Data Collection and Preprocessing:

- Gather a dataset containing symptom-based information for various medical conditions.
- The data set used for this work has 42 disease label which is based on 132 symptoms. Data set can be downloaded from this link (https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning/data)
- The classes includes fungal infection, hepatitis C, hepatitis E, Alcoholic hepatitis, Tuberculosis, common cold, pneumonia, dimorphic haemorrhoids(pile), heart attack, Varicose veins, hypothyroidism, hyperthyroidism, hypoglycemia, osteoarthiristis, arthritis (vertigo), paroymsal positional vertigo, acne, urinary tract infection, psoriasis, hepatitis D, hepatitis B, allergy, hepatitis A, GERD, chronic cholestasis, drug reaction, peptic ulcer disease, AIDS, diabetes, gastroenteritis, bronchial asthma, hypertension, migraine, cervical spondylosis, paralysis(brain hemorrhage), jaundice, malaria, chicken pox, dengue, typhoid, impetigo
- Preprocess the data to handle missing values, encode categorical variables.

2) Feature Selection:

- Explore advanced feature selection techniques such as Recursive Feature Elimination (RFE), Mutual Information, Information Gain.
- Evaluate the relevance of features to identify the most informative subset for disease prediction.
- Used combined knowledge of all these methods to identify

the top features

3) Machine Learning Model Selection:

- Choose appropriate machine learning algorithms for multi-label classification such as Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), etc.
- Assess the suitability of each algorithm based on performance metrics and computational efficiency.

4) Hyperparameter Optimization:

- Optimize the hyperparameters of selected machine learning models using techniques like grid search and cross-validation.
- Fine-tune the parameters to maximize model performance while avoiding overfitting.

5) Model Training and Evaluation:

- Train the machine learning models on the training dataset using the selected features and optimized hyperparameters.
- Evaluate the trained models on a separate validation dataset to assess their performance in terms of accuracy, precision, recall.

6) Results Interpretation and Discussion:

- Interpret the results obtained from the experiments and discuss their implications for disease prediction and diagnosis.
- Identify strengths, limitations, and potential areas for future research.

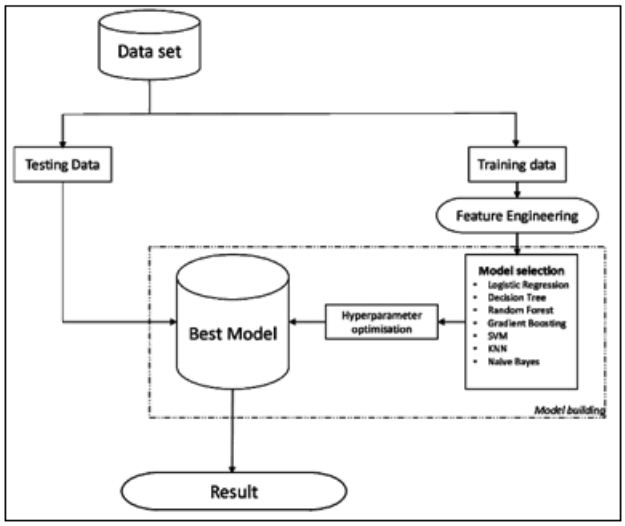

Figure 1: Proposed methodology diagram

Recursive Feature Elimination (RFE):

Recursive Feature Elimination is a feature selection technique commonly employed in machine learning. It operates by recursively removing attributes and building a model on those attributes that remain. This process continues until the specified number of features is reached. RFE evaluates the importance of each feature by considering the impact of its removal on the performance of the model. By iteratively pruning less important features, RFE helps to identify the most relevant subset of features for modeling, thereby enhancing model efficiency and interpretability.

Let X be the feature matrix with dimensions m×n, where mm is the number of samples and nn is the number of features. Let y be the target vector with dimensions m×1. The goal of RFE is to recursively eliminate features to identify the subset that optimally contributes to model performance.

The algorithm iterates as follows:

- Train a machine learning model on the dataset (X,y).
- Rank the features based on their importance scores derived from the trained model.
- Remove the least important feature.
- Repeat the process until the desired number of features is obtained.

Mutual Information:

Mutual Information is a measure of the statistical dependence between two variables. In the context of feature selection, Mutual Information assesses the amount of information obtained about one variable through the other variable. It quantifies the degree of association between features and the target variable, indicating how much knowing one feature reduces uncertainty about the other. By calculating Mutual Information scores for each feature, it identifies the most informative attributes for predictive modeling, facilitating the selection of relevant features that contribute significantly to the predictive performance of the model.

- Mutual Information between two random variables Xand Y is defined as:
  $I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log (p(x, y) / p(x)p(y))$
  where p (x, y) is the joint probability distribution function of X and Y, and p (x) and p (y)are the marginal probability distribution functions of Xand Y, respectively.
- For feature selection, the Mutual Information between each feature and the target variable y is computed, and features with high Mutual Information scores are selected as they provide more information about the target.

Information Gain:

Information Gain is a concept borrowed from information theory and is commonly used in decision tree algorithms for feature selection. It measures the reduction in entropy or uncertainty achieved by splitting a dataset based on a particular feature. Higher Information Gain implies that a feature provides more discriminatory power in partitioning the data into distinct classes. Features with higher Information Gain are considered more relevant for classification tasks as they contribute more towards differentiating between classes, thereby improving the overall accuracy of the predictive model.

- Information Gain is calculated based on the concept of entropy. Given a dataset DD with classes {C1, C2 ,…., Ck}, the entropy H(D) is defined as:
  $H(D) = -\sum_{i=1} p(Ci) \log_2 p(Ci)$
  where p(Ci)) is the probability of class Ci occurring in the dataset.
- The Information Gain of a feature A with respect to the dataset D is given by:
  $IG(D, A) = H(D) - \sum_{v \in Values(A)} |D_v| / |D| * H(D_v)$
  where Values(A) represents the possible values of feature A, D_v is the subset of D for which feature A has value v, and |Dv| is the number of instances in Dv.

- Features with higher Information Gain are preferred as they provide more discriminatory power in classifying instances.

LASSO (Least Absolute Shrinkage and Selection Operator):
LASSO is a regularization technique widely utilized in regression analysis and feature selection. It penalizes the absolute size of the regression coefficients, forcing some coefficients to shrink to zero, effectively performing feature selection by eliminating irrelevant variables. LASSO encourages sparse solutions by imposing a constraint on the sum of the absolute values of the coefficients, thereby promoting model simplicity and interpretability. By shrinking certain coefficients to zero, LASSO identifies and retains only the most important features while discarding less influential ones, facilitating the construction of parsimonious and effective predictive models.

The ||L1|| term is the L1 norm penalty, which encourages sparsity by penalizing the absolute size of the coefficients. As λ increases, more coefficients are driven to zero, leading to feature selection. Features with non-zero coefficients in the optimized model are retained, while features with zero coefficients are discarded

4. Result and Discussion

The training dataset comprises 4962 observations from patients, ensuring a robust representation of medical conditions. Notably, the dataset exhibits a balanced distribution across all 42 classes and is devoid of any missing values, ensuring the integrity of the analysis.

The distribution of class labels across the dataset, portraying the prevalence of various medical conditions. This visualization provides valuable insights into the relative frequency of each class, aiding in understanding the dataset's composition. In anticipation of identifying the most influential features, a plot showcasing the top 10 symptoms is presented (Figure 3). This visual aid serves as a reference for subsequent feature selection endeavours, providing additional evidence to support the findings

.

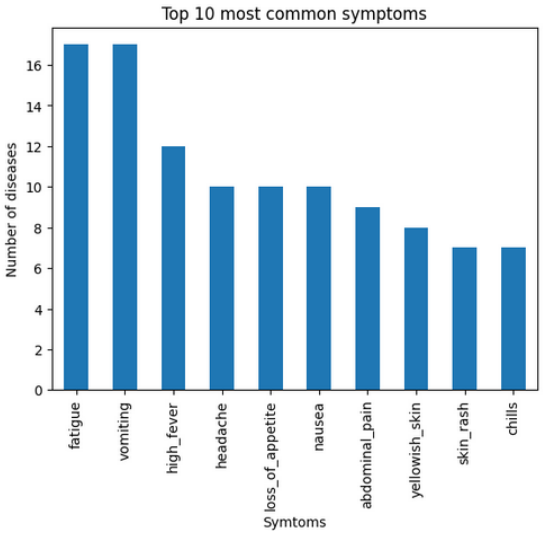

Figure 2: Plot of top 10 features

"Building upon these observations, Recursive Feature Elimination (RFE) is employed to determine an optimal feature subset. Figure 4 depicts the relationship between the number of features and model accuracy, highlighting a range wherein the model achieves optimal performance

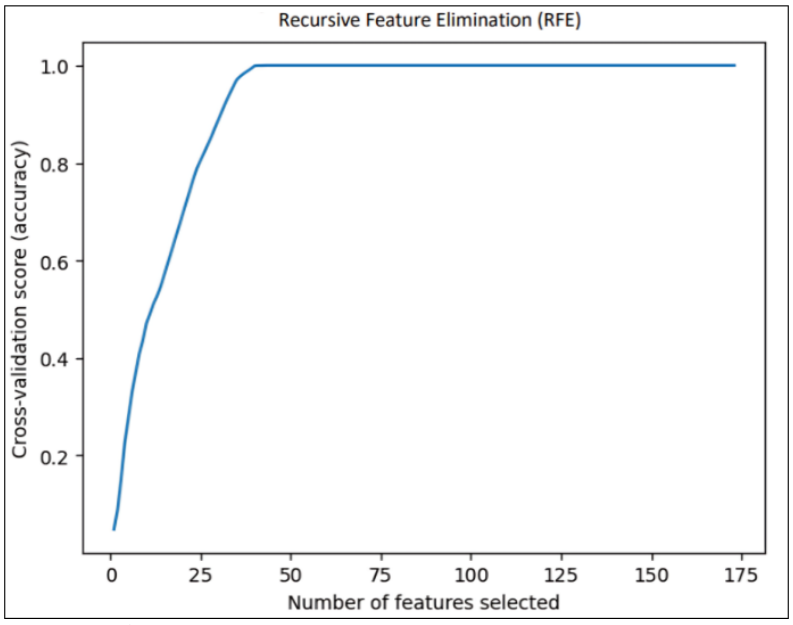

Figure 3: Plot of optimal number of feature based on RFE

To assess baseline model performance prior to feature engineering, a basic machine learning model is constructed (Figure 5). Results indicate near-perfect accuracy when utilizing the entire feature set, suggesting potential overfitting and impracticality for real-world applications

After this a basic machine learning is model is build to have an idea for the accuracy before feature engineering Figure.5. We can observe that most of the model gave perfect accuracy of course by using 132 feature. This means the model is over

fitting and also the training data requirement is unreliable for real world application.

	Model	Training Accuracy	Validation Accuracy
0	Logistic Regression	1.0	1.00000
1	Decision Tree	1.0	0.97619
2	Random Forest	1.0	0.97619
3	Gradient Boosting	1.0	0.97619
4	SVM	1.0	1.00000
5	K-Nearest Neighbors	1.0	1.00000
6	Naive Bayes	1.0	1.00000

Figure 4: Accuracy before feature engineering

Following thorough data visualization and analysis, feature engineering techniques are applied based on insights from Recursive Feature Elimination, Mutual Information Gain, Information Gain Score, and Lasso selector. Twenty-five features, including abdominal pain, chest pain, chills, and others, are identified as the most influential contributors to the classification task abdominal_pain, chest_pain, chills, constipation, diarrhoea, dizziness, family_history, fatigue, headache, high_fever, irritability, itching, joint_pain, loss_of_appetite, loss_of_balance, malaise, muscle_pain, muscle_weakness, nausea, skin_rash, stomach_pain, sweating, vomiting, weight_loss, yellowing_of_eyes

We can confirm these feature by our previous observation from Figure 3, 4. All the top ten features from the plot are there in this selected features list and 25 is also estimated as a optimal number from Figure 4.

Subsequent evaluation of the model post-feature engineering demonstrates a notable improvement in accuracy, despite a significant reduction in the number of features by 81%. This enhancement underscores the efficacy of feature selection techniques in refining model performance and mitigating overfitting (Figure 6).

	Model	Training Accuracy	Validation Accuracy
0	Logistic Regression	0.902439	0.928571
1	Decision Tree	0.902439	0.928571
2	Random Forest	0.902439	0.928571
3	Gradient Boosting	0.902439	0.928571
4	SVM	0.902439	0.928571
5	K-Nearest Neighbors	0.901220	0.928571
6	Naive Bayes	0.882927	0.904762

Figure 5: Accuracy after feature engineering

Furthermore, comprehensive performance metrics including precision, recall, F1-score, and hamming loss, along with a classification report, are provided to further evaluate the model's effectiveness in disease prediction and diagnosis.

```
Average Precision (micro): 0.8943089430894309
Average Recall (micro): 0.8943089430894309
Average F1-score (micro): 0.8943089430894309
Average Hamming Loss: 0.1056910569105691
```

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       0.50      1.00      0.67         1
           2       0.00      0.00      0.00         1
           3       1.00      1.00      1.00         1
           4       1.00      1.00      1.00         1
           5       1.00      1.00      1.00         1
           6       1.00      1.00      1.00         1
           7       1.00      1.00      1.00         1
           8       1.00      1.00      1.00         1
           9       1.00      1.00      1.00         1
          10       1.00      1.00      1.00         1
          11       1.00      1.00      1.00         1
          12       1.00      1.00      1.00         1
          13       1.00      1.00      1.00         1
          14       1.00      1.00      1.00         1
          15       1.00      1.00      1.00         2
          16       1.00      1.00      1.00         1
          17       1.00      1.00      1.00         1
          18       1.00      1.00      1.00         1
          19       1.00      1.00      1.00         1
          20       1.00      1.00      1.00         1
          21       1.00      1.00      1.00         1
          22       1.00      1.00      1.00         1
          23       1.00      1.00      1.00         1
          24       1.00      1.00      1.00         1
          25       1.00      1.00      1.00         1
          26       1.00      1.00      1.00         1
          27       0.00      0.00      0.00         1
          28       1.00      1.00      1.00         1
          29       1.00      1.00      1.00         1
          30       1.00      1.00      1.00         1
          31       0.00      0.00      0.00         1
          32       1.00      1.00      1.00         1
          33       1.00      1.00      1.00         1
          34       1.00      1.00      1.00         1
          35       0.50      1.00      0.67         1
          36       1.00      1.00      1.00         1
          37       1.00      1.00      1.00         1
          38       0.50      1.00      0.67         1
          39       1.00      1.00      1.00         1
          40       1.00      1.00      1.00         1

    accuracy                           0.93        42
   macro avg       0.89      0.93      0.90        42
weighted avg       0.89      0.93      0.90        42
```

Figure 6: Classification report of the final model

5. Conclusion

Our results demonstrate the effectiveness of the proposed approach in improving the accuracy and reliability of disease classification models. By incorporating feature engineering and hyperparameter optimization, we achieved a more practical and deployable machine learning model, reducing the risk of overfitting and enhancing the model's generalization capabilities. The findings of this research have significant implications for early disease recognition, improved patient care, and public health interventions. The developed model can serve as a valuable tool for healthcare professionals, aiding in the timely diagnosis and treatment of various medical conditions. Moving forward, further validation and refinement of the model on diverse datasets and real-world clinical settings are warranted to ensure its effectiveness and applicability in practical healthcare scenarios. Additionally, ongoing advancements in machine learning and data analysis techniques offer promising opportunities for continuous improvement and innovation in medical diagnosis and patient care. This proposed methodology can be implemented in a hospital's website for quick initial diagnosis of disease. This work has significantly reduced the number of questions need to be asked by 81% that is in place of 132 questions only 25 needed to be asked. The model has shown an accuracy of 92.85 with average F1 score of 89.4.

References

[1] Bhanuteja Talasila, Saipoornachand Kolli, Kilaru Venkata Narendra Kumar and Poonati Anudeep, "Symptoms Based multiple Disease Prediction Model using Machine Learning Approach", International

Journal of Innovative Technology and Exploring Engineering, August 2021.
[2] Ibrahim I. and Abdulazeez, A., 2021. The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, *2*(01), pp.10-19.
[3] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang,"Disease Prediction by Machine Over Learning Over Big Data From Healthcare Communities," in IEEE Access, vol.5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446
[4] Parvez Ahmad, Saqib Qamar, Syed Qasim Afser Rizvi. "Techniques of Data Mining in Healthcare: A Review." International Journal of Computer Applications, Volume 120 – No.15, June 2017.
[5] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, William Hersh. "Knowledge Management, Data Mining, and Text Mining in Medical Informatics."
[6] Ibrahim Mahmood and Adnan Moshin Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases", Journal of Applied Science and Technology Trends (JASTT), March 2021.
[7] Divya Tomar, Sonali Agarwal. "A survey on data mining approaches for healthcare." International Journal of Bio-Science and Bio-Technology, Vol. No.5, pp. 241-266, 2017.
[8] Mohammed Abdul Khalid, Sateesh Kumar Pradhan, G.N. Dash, F.A. Mazarbhuiya. "A survey of data mining techniques on medical data for finding temporally frequent diseases." International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 12, December 2018.
[9] S.D. Gheware, A.S. Kejkar, S.M. Tondare. "Data Mining: Task, Tools, Techniques and Applications." International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 10, October 2017.
[10] Yongjian Fu. "Data Mining: Tasks, Techniques and Applications."
[11] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. "Introduction to Data Mining", Addison Wesley, 2017.
[12] G. Beller. "The rising cost of health care in the United States: is it making the United States globally noncompetitive?"
[13] Sneha Grampurohit and Chetan Sagamal, "Disease Prediction using Machine Learning Algorithms", IEEE International Conference for Emerging Technology (INCET), August 2020..
[14] Gosain, A., Kumar, A. "Analysis of health care data using different data mining techniques." Intelligent Agent & Multi-Agent Systems, 2009, International Conference on, 1-6, July 22-24, 2018.
[15] Dr. M.H. Dunham. "Data Mining: Introductory and Advanced Topics."
[16] A.S. Elmaghraby et al. "Data Mining from multimedia patient records."
[17] Nada Lavrac, Blaž Zupan. "Data Mining in Medicine" in Data Mining and Knowledge Discovery Handbook.
[18] Soni J, Ansari U, Sharma D. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction." International Journal of Computer Applications, Volume 17– No.8, March 2018.
[19] Naren Ramakrishnan, David Hanauer, Benjamin J. Keller. "Mining Electronic Health Records." IEEE Computer, 43(10): 77-81, 2018.
[20] O. Mary K, Mat. "Applications of Data Mining Techniques to Healthcare Data." Infection Control and Hospital Epidemiology, August 2017.
[21] Hian Chye K, Gerald T. "Data mining applications in healthcare." Journal of healthcare information management: JHIM, 19 (2): 64-72, 2016.
[22] A. Milley. "Healthcare and data mining." Health Management Technology, Vol. 21, No. 8, pp. 44-47, 2017.
[23] Gaynes R, Richards C, Edwards J, et al. "Feeding back surveillance data to prevent hospital-acquired infections." Emerg Infect Dis 2001; 7:2 95298, 2017.
[24] Brosette SE, Spragre AP, Jones WT, Moser SA. "A data mining system for infection control surveillance." Methods Inf Med, 39: 303-310, 2018.
[25] M. Ridinger. "American Healthways uses SAS to improve patient care." DM Review, Vol. 12, No.139, 2018.

Author Profile

Suraj Kumar is a Lead Data Scientist where my passion for data science powers product innovation and impactful outcomes. My journey has taken me through roles at Walmart, Intuit, and Realtor.com, where I've specialized in experimentation and product analytics. At Walmart, I delved into consumer behaviors to refine shopping experiences, boosting engagement and sales. At Intuit, I used data to tailor financial tools, making complex decisions simpler for users. Realtor.com saw me merging real estate trends with digital preferences to enhance property searches. My mantra is simple: listen to the data, iterate based on insights, and always aim for products that are not just better, but are also more intuitive and aligned with user needs.

Kukku Youseff is a post graduate student in data science doing research in deep learning and machine learning under the guidance of first author.