Abstract

**Purpose:** Health recommenders act as important decision support systems, aiding patients and medical professionals in taking actions that lead to patients’ well-being. These systems extract the information which may be of particular relevance to the end-user, helping them in making appropriate decisions. The present study proposes a feature recommender, as a part of a disease management system, that identifies and recommends the most important risk factors for an illness.

**Methods:** A novel mutual information and ensemble-based feature ranking approach for identifying critical risk factors in healthcare prognosis is proposed.

**Results:** To establish the effectiveness of the proposed method, experiments have been conducted on four benchmark datasets of diverse diseases (clear cell renal cell carcinoma (ccRCC), chronic kidney disease, Indian liver patient, and cervical
cancer risk factors). The performance of the proposed recommender is compared with four state-of-the-art methods using recommender systems’ performance metrics like average precision@K, precision@K, recall@K, F1@K, reciprocal rank@K. The method is able to recommend all relevant critical risk factors for ccRCC. It also attains a higher accuracy (96.6% and 98.6% using support vector machine and neural network, respectively) for ccRCC staging with a reduced feature set as compared to existing methods. Moreover, the top two features recommended using the proposed method with ccRCC, viz. size of tumor and metastasis status, are medically validated from the existing TNM system. Results are also found to be superior for the other three datasets.

**Conclusion:** The proposed recommender can identify and recommend risk factors that have the most discriminating power for detecting diseases.

**Keywords:** Critical Risk Factors, Feature Recommender, Healthcare Prognosis, Mutual Information, Disease Detection

1 Introduction

In the healthcare domain, the vast amount of clinical data has created an overload of information in terms of medical tests, drugs, and treatment options. This overload leads to indecisiveness in various health-related scenarios both for medical professionals and patients. In such a situation, the application of recommender systems, as a part of a disease management system, lead to accurate and timely patient-related decisions. The healthcare recommender system (HRS) can help the end-users as well as the medical professionals in multiple domains. Example areas include recommendations for diet, drug, and treatment as well as feature recommendations to assist the development of machine learning-based healthcare prognostic systems.

In this article, an automated feature recommender using mutual information with an ensemble strategy is proposed to identify clinically relevant features that can help in disease prognosis. Cervical cancer, chronic kidney disease, liver disease, and clear cell renal cell carcinoma, with high fatality rates, are selected to show the effectiveness of this study. Eight feature selection methods including two wrapper, five filter, and one embedded technique, are used initially to rank the features. From the ranking information obtained by these eight methods, an ensemble approach is exploited to obtain the final positional feature recommendation table. This positional feature recommendation table contains the features in a ranked manner for further processing. The performance of the proposed method is evaluated using various performance metrics used for recommender systems like average precision@K, precision@K, recall@K, F1@K, and reciprocal rank@K.

The contribution of this work can be highlighted as:

- Development of a health recommender system that helps in identifying the critical risk factors for healthcare prognosis.
An unique ensemble strategy is developed to rank features for the recommendation task. This strategy leverages the strengths of eight different feature selection algorithms, potentially leading to a more robust and informative selection of risk factors compared to using the individual algorithms.

The effectiveness of the proposed method is shown on four standard disease datasets including a detailed demonstration for stage progression of clear cell renal cell carcinoma.

Two supervised classification models are trained using the identified risk factors. The performance of these models is assessed to verify the effectiveness of the recommender system.

The rest of the article is organized as follows. Preliminaries for this study and the proposed method is discussed in Section 2. Section 3 includes the datasets, the performance metrics used, the experimental results obtained along with its interpretation. Section 4 concludes the article with closing remarks and the future directions of this research.

2 Materials and Methods

This section describes the preliminaries for this study and explains the proposed method.

2.1 Healthcare Recommender Systems

With the huge amount of information available online, recommender system acts as a decision support system providing recommendations of items/services which could be useful to the user. This may be in the form of what items to buy, what music to listen, what treatment to undergo [1], etc. Health recommender system is a special area of application in recommender systems where medical items, knowledge, or services are recommended to the end users (patients, medical professionals, etc.)

Health recommender systems try to build constructive knowledge out of diverse medical information and combine it with personal health data to make recommendations that can help in healthcare-related decision making as shown in Figure 1. For the patient, this decision support can be in the form of diet/treatment/lifestyle/drug/doctor/hospital recommendation. For the medical professional, this decision support can be in the form of identification of the risk factors of a particular disease, region of interest detection in medical images as well as the recommendation of drug/treatment options.

2.2 Related Works

Proper feature selection not only improves the performance but also plays a key factor in bringing explainability to the system. Being able to explain a recommendation or a classification and attributing it to certain features becomes easier when the number of features is less. Thus, feature selection remains an active area of research in the healthcare domain. In [2, 3], the authors performed a comprehensive survey on feature selection for healthcare domain. Multiple studies [4–6] exist that utilised soft
computing-based optimization techniques like genetic algorithm, particle swarm optimization, ant colony optimization for feature selection, and pattern classification. In [7], wrapper-based feature selection is done using seven common feature selection methods and it is concluded from the experiment that Bayesian network with top 15 features is the most suitable prediction model. However, in [7], the medical relevance of the selected features is not sufficiently analyzed. Filter and wrapper-based feature selections have been used in many studies [8, 9] for tasks like activity recognition, clinical decision making, classification of medical conditions, etc.

HRS-based decision support system is also becoming an important part of healthcare prognosis. These systems aid better decision-making for both patients as well as healthcare professionals. Both content-based as well as collaborative methods have been explored for the design of HRS [10, 11]. HRS has been proposed for recommending the best doctors [12], suitable drugs [13], personalized treatment option [14], proper diet [15], managing COVID-19 crisis [16] and multiple other areas.

Several works exist in the literature that recommend attributes or deal with feature ranking methods in various domains. [17] presented the use of multi-criteria decision-making for creating a decision support system for e-commerce products. [18] uses various collaborative and content-based recommendation systems for Turkish movie recommendation and also shows how feature selection improves the recommendation results. [19] builds a recommender system for ranking the most important attributes for breast cancer prognosis using the correlation criterion. [20] proposes a feature recommender using multi-criteria decision-making and genetic algorithm for a cervical cancer dataset. [21] builds a model for the recommendation of a feature selection algorithm using a meta-learning objective.

The work presented in this article explores the area of identification and recommendation of clinical risk factors for healthcare prognosis. A feature recommender system has been developed for clear cell renal cell carcinoma staging, then it is generalized for other benchmark datasets. Unlike the existing methods cited above, this article uses an ensemble of multiple feature selection techniques for feature recommendation.
2.3 Proposed Method

Eight feature selection methods have been used for constructing the final positional feature recommendation table, from which the most relevant features may be recommended as shown in Figure 2. As wrapper methods, Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are used because of their implementation simplicity and efficiency in producing a set of features that maximize classification accuracy [22]. Among filter methods, three mutual information (MI)-based (MIFS [23], joint mutual information (JMI) [24] and normalized mutual information (NMI) [25]) have been used in this work. The other two filter methods used are Chi-square and F-score [26]. As an embedded approach of feature selection, random forest [27] is used which tends to select a good subset of features that can avoid over-fitting of data without compromising classification accuracy.

The aim of the proposed article is to identify the features that provide insight into the internal characteristics of data and can also maximize classification accuracy without over-fitting. Such features are expected to be highly ranked by the majority of filter, wrapper, and embedded approaches used in the experiments. Keeping this in mind, a positional table $T$ has been implemented which takes into account the rank of features obtained through the previously mentioned feature selection methods.

Let the number of features in the dataset be denoted as, $NF$, and the number of feature selection techniques used is denoted as, $NT$. In our experiment, $NT$ is eight. Each location of $T$ can be defined as, $T(i, j) = \text{feature index with rank } i \text{ obtained using feature selection method } j (i = 1 \text{ to } NF, j = 1 \text{ to } NT)$. Hence, in the table each row corresponds to the rank of the feature and columns indicate feature selection methods. From this table, the importance of any feature is obtained by observing its position. If, for a specific feature all of its positions occur near the top of this table then it can be concluded that the said feature is significant. If, for one particular row ($r$), a single feature ($f$) occurs in the majority of its column positions then it will be selected for inclusion with rank $r$ in the final set. But, if such a single feature is not found then a feature having a higher correlation with the ground truth (i.e., class labels of the
dataset) will be selected. This correlation can be measured from the MI-based ranking criterion given below [28]:

\[ \text{MIrank}(f) = \sum_{k=1}^{|C|} p(c_k) \text{MI}(c_k, f), \]  

(1)

where, \(|C|\) is the total number of classes, \(p(c_k)\) is the probability of the class \(c_k\), \(\text{MI}(c_k, f)\) is the mutual information between class \(c_k\) and feature \(f\). Hence, in the absence of a single feature occurring in row \(r\) for the majority of times, the feature having the highest MI value with the class labels will be selected with rank \(r\). If all the features in row \(r\) have already been included in \(S\), no new feature can be selected as rank \(r\) for that row. In such case, the next selected feature for row \((r + 1)\) will be promoted to rank \(r\). This will not impose any problem in the feature selection process as it does not violate the ordering of the selection of features.

2.3.1 Ensemble-based Feature Recommender

The proposed feature ensemble approach is formally described in Algorithm 1. The process is initialized with \(S = \emptyset\) and \(R = \) Original set of features in the dataset. \(\text{Occ}[]\) is an array that is used for storing the number of occurrences of each feature encountered in a single row of positional table \(T\) i.e., \(\text{Occ}[f]\) will give the number of occurrences of feature \(f\) in a single row. Evidently, \(\text{Occ}[f] \leq NT\). \(\text{Rank}[]\) array stores the final rank of features after the execution of the selection process, \(\text{Rank}[i]\) will indicate the \(i^{th}\) ranked feature (\(i = 1 \text{ to } NF\)).

Algorithm 1: Pseudocode of the Proposed Algorithm

1. \(S = \emptyset\) //Final set of features, initially empty
2. \(R = \) Original set of features in the dataset
3. Construct a positional table \(T\) so that \(T(i, j)\) contains feature index with rank \(i\) using feature selection method \(j\) (\(i = 1 \text{ to } NF, j = 1 \text{ to } NT\))
4. For \(i = 1 \text{ to } NF\) do
   (a) Construct a set \(F\) that contains different features with rank \(i\) (as obtained from the \(i^{th}\) row of \(T\))
   (b) Calculate the number of occurrences of each feature \(f \in F\) and store the value in \(\text{Occ}[f]\)
   (c) If (for a specific feature \(sf\), \(\text{Occ}[sf]\) is higher than other features in \(F\)) AND \((sf \notin S)\)
       \(\bullet S = S \cup sf\) //Updating final set of selected features by including currently selected feature, sf
       \(\bullet R = R - sf\) //Updating remaining set of features by excluding currently selected feature, sf
   (d) ElseIf (no such single feature is found) then
       \(\bullet S = S \cup \{sf\}\)
       \(\bullet R = R - \{sf\}\)
Table 1 Brief description of benchmark datasets

| Dataset                                      | #Features | #Classes | #Samples |
|----------------------------------------------|-----------|----------|----------|
| Chronic Kidney Disease [29]                  | 24        | 2        | 400      |
| Indian Liver Patient [30]                    | 10        | 2        | 583      |
| Cervical Cancer Risk Factors [31]            | 36        | 2        | 858      |
| Clear Cell Renal Cell Carcinoma (ccRCC)      | 29        | 4        | 416      |

EndIf
(e) \( Rank[i] = sf \) //finally \( sf \) is selected as the \( i^{th} \) ranked feature
EndFor

The above pseudocode provides the sequence of operations to be performed in order to obtain the final rank of features received from the \( Rank[] \) array. The 1\(^{st}\) location of \( Rank[] \) contains the feature with the highest rank, the 2\(^{nd}\) location contains the feature with the 2\(^{nd}\) highest rank, and so on. In order to recommend \( d \) higher ranked features (\( d < NF \)), the first \( d \) features from the \( Rank[] \) array are to be kept in \( S \) and other features are to be discarded.

3 Results and Discussion

This section explains the datasets and illustrates the results obtained with their interpretation.

3.1 Datasets

Experiments have been conducted on four datasets: chronic kidney disease [29], Indian liver patient [30], cervical cancer risk factors [31], and clear cell renal cell carcinoma. A brief description of these datasets is given in Table 1. Our custom aggregated dataset on ccRCC is obtained from The Cancer Genome Atlas data collections [32, 33]. The dataset is a combination of clinical, histopathological, demographical, and follow-up for staging of renal cell carcinoma. The dataset consists of twenty-nine features and four classes representing four different stages of renal cancer. A description of these features along with their feature numbers are given in Table 2. For a detailed illustration, the steps of the proposed method are shown on the ccRCC dataset.

3.2 Performance Metrics

The performance metrics used in this study are precision@K, recall@K, average precision@K, F1@K and reciprocal rank@K [34]. These metrics assess the quality of recommendation given by the proposed system, where \( K \) is the number of recommended items specified by the user. For this article, items refer to the relevant features for prognosis.

3.3 Positional Feature Recommendation Table

The positional table obtained for clear cell renal cell carcinoma dataset using Algorithm 1 (as explained in Section 4.1) is shown in Table 3. From this table, it is evident
| Feature Number | Name                                      | Description                                              | Type          |
|----------------|-------------------------------------------|----------------------------------------------------------|---------------|
| 0              | tumor grade                              | tumor grade G score                                      | Categorical   |
| 1              | laterality                                | location of ccRCC kidney                                 | Binary        |
| 2              | gender                                    | patient’s gender                                         | Binary        |
| 3              | history other malignancy                  | history of other tumor disease                            | Binary        |
| 4              | history neoadjuvant treatment             | history of some chemical therapy                         | Binary        |
| 5              | lymph nodes examined count                | problem in lymph node                                    | Continuous    |
| 6              | lymph nodes examined positive             | number of cancer nodes                                   | Continuous    |
| 7              | ajcc tumor pathologic pt                  | AJCC T phase indicator                                   | Continuous    |
| 8              | ajcc nodes pathologic pn                  | AJCC N phase indicator                                   | Categorical   |
| 9              | ajcc metastasis pathologic pm             | AJCC M phase indicator                                   | Binary        |
| 10             | serum calcium level                       | calcium level in blood                                   | Categorical   |
| 11             | haemoglobin level                         | haemoglobin level in blood                               | Categorical   |
| 12             | platelet count                            | number of platelets in blood 1mm²                        | Categorical   |
| 13             | white cell count                          | number of white cells in blood 1mm²                      | Categorical   |
| 14             | percent stromal cells                     | percentage of stromal cells                              | Continuous    |
| 15             | percent tumor cell                        | percentage of tumor cells                                | Continuous    |
| 16             | intermediate dimension                    | dimension related parameter                              | Continuous    |
| 17             | longest dimension                         | dimension related parameter                              | Continuous    |
| 18             | shortest dimension                        | dimension related parameter                              | Continuous    |
| 19             | percent necrosis                          | tumor necrosis percentage                                | Continuous    |
| 20             | percent normal cells                      | percentage normal cells                                  | Continuous    |
| 21             | percent tumor nuclei                      | percentage tumor nuclei                                  | Continuous    |
| 22             | birth days to                             | lifetime in days                                         | Continuous    |
| 23             | initial pathologic dx year                | initial cancer notification                              | Continuous    |
| 24             | lymph nodes examined                     | number of tumors in lymph node                           | Categorical   |
| 25             | vital status                              | patient’s survival indicator                             | Categorical   |
| 26             | last contact days to                      | number of days to last advise/contact                     | Continuous    |
| 27             | tumor status                              | another kind of tumor except ccRCC                       | Binary        |
| 28             | tobacco smoking history indicator         | patient’s smoking indicator                               | Categorical   |

that feature numbers 7 and 9 are the two most important ones as they are selected by most of the feature selection methods as top-ranked features. Therefore, they are ranked as 1 and 2, respectively, in the final rank. For row numbers: 1, 2, 4, 5, 6, 10, 13, 14, 16, 18, 20, 21, 27, 29 in Table 3, we observe the occurrence of single features (feature numbers 7, 9, 0, 27, 1, 8, 18, 19, 13, 12, 23, 10, 6, 26, respectively as given in column 3 of Table 4) in the majority. For other rows, no such single feature is found and the final selection is done using Equation 1. For each feature, MI-based rank is shown in the second column of Table 4. Also, for a particular row, already included features in the final set $S$ must not be considered for selection to avoid duplication of features in $S$. It can be observed that for row 23, no feature can be selected because all the features present in row 23 are already included in $S$. Therefore, the next selected feature in row 24 (feature 16) can be considered as rank 23. Note that, it will not violate the ordering of the feature selection process. Once the selection is over, we get 28 features in $S$. Feature number 26 is the one not selected so far. So, it has been included as the last feature (with rank 29) in $S$. In this way a final ranking of features is obtained by the proposed method as shown in the last column of Table 4.

According to the American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC), there exists a TNM system [35] to classify the four stages of renal cell carcinoma. In this system, feature numbers: 7(T - tumor size),
Table 3: Ranking of features obtained based on 8 feature selection methods. The numbers in columns 2-9 indicate the feature numbers as shown in Table 2.

| Feature Rank | Feature Selection Method |
|--------------|--------------------------|
|              | Random Forest  | SFS  | SBS  | MIFS [23] | Chi-Squared | F-score | JMI [24] | NMI [25] |
| 1            | 7              | 7    | 7    | 22         | 26           | 9       | 22       | 7        |
| 2            | 9              | 9    | 9    | 9          | 7             | 7       | 7        | 9        |
| 3            | 27             | 6    | 0    | 7          | 22            | 27      | 26       | 22       |
| 4            | 22             | 0    | 15   | 6          | 9             | 0       | 9        | 26       |
| 5            | 26             | 15   | 6    | 4          | 5             | 25      | 27       | 27       |
| 6            | 0              | 1    | 1    | 20         | 27            | 26      | 0        | 0        |
| 7            | 17             | 2    | 2    | 3          | 14            | 13      | 17       | 25       |
| 8            | 14             | 3    | 3    | 27         | 19            | 24      | 14       | 5        |
| 9            | 16             | 4    | 4    | 28         | 25            | 3       | 25       | 17       |
| 10           | 18             | 8    | 8    | 24         | 20            | 17      | 16       | 14       |
| 11           | 23             | 10   | 11   | 2          | 24            | 18      | 5        | 19       |
| 12           | 5              | 11   | 12   | 12         | 0             | 5       | 15       | 15       |
| 13           | 25             | 5    | 18   | 1          | 13            | 11      | 18       | 16       |
| 14           | 15             | 17   | 19   | 19         | 6             | 8       | 23       | 13       |
| 15           | 21             | 20   | 22   | 11         | 28            | 4       | 19       | 18       |
| 16           | 13             | 22   | 10   | 13         | 3             | 28      | 21       | 24       |
| 17           | 24             | 13   | 17   | 25         | 4             | 6       | 13       | 20       |
| 18           | 12             | 26   | 20   | 10         | 8             | 12      | 24       | 12       |
| 19           | 3              | 19   | 21   | 8          | 17            | 1       | 8        | 6        |
| 20           | 8              | 21   | 26   | 0          | 23            | 23      | 12       | 3        |
| 21           | 10             | 12   | 14   | 21         | 15            | 2       | 10       | 8        |
| 22           | 11             | 14   | 24   | 18         | 12            | 15      | 20       | 21       |
| 23           | 1              | 18   | 23   | 5          | 1             | 22      | 3        | 23       |
| 24           | 28             | 16   | 25   | 23         | 11            | 14      | 6        | 10       |
| 25           | 2              | 23   | 27   | 16         | 18            | 19      | 11       | 4        |
| 26           | 19             | 24   | 28   | 15         | 2             | 20      | 4        | 11       |
| 27           | 4              | 28   | 16   | 17         | 16            | 16      | 28       | 28       |
| 28           | 6              | 25   | 13   | 14         | 21            | 21      | 1        | 1        |
| 29           | 20             | 27   | 5    | 26         | 10            | 10      | 2        | 2        |

8(N - number of lymph nodes examined positive) and 9(M - metastasis status) are the determining factors of the cancer stage. The other relevant risk factors for renal cancer staging as found from literature are feature numbers: 0(tumor grade), 27 (presence of another kind of tumor), 1(location of ccRCC kidney), 17(longest dimension of the tumor) and 5(number of lymph nodes examined) [36–39].

If we consider the top 11 recommendations (K = 11), the proposed method is able to identify all 8 risk factors, depicting better performance with respect to the individual feature selection algorithms used, as shown in Table 5.

3.4 Comparison with State-of-the-art Methods on ccRCC dataset

The proposed method of feature recommendation has been compared with four state-of-the-art (SOTA) works published between 2016 and 2023 (Baczkiewicz et al. [17], Cataltepe et al. [18], Kanimozhi et al. [18], Kuans et al. [20]) on the ccRCC dataset. The results are shown in the Table 6 and in the form of a graph in Figure 3. Reciprocal rank (RR) is the same for all K as the rank of the first relevant feature is the same for all K.
Table 4  Ranking of features obtained based on mutual information (MI) method and our proposed feature selection method. The numbers in columns 2 and 3 indicate the feature numbers as shown in Table 2

| Feature rank | Mutual Information Based Method | Proposed Method |
|--------------|---------------------------------|-----------------|
| 1            | 22                              | 7               |
| 2            | 7                               | 9               |
| 3            | 26                              | 22              |
| 4            | 9                               | 0               |
| 5            | 27                              | 27              |
| 6            | 0                               | 1               |
| 7            | 17                              | 17              |
| 8            | 14                              | 14              |
| 9            | 25                              | 25              |
| 10           | 16                              | 8               |
| 11           | 5                               | 5               |
| 12           | 15                              | 15              |
| 13           | 18                              | 18              |
| 14           | 23                              | 19              |
| 15           | 19                              | 21              |
| 16           | 21                              | 13              |
| 17           | 13                              | 24              |
| 18           | 24                              | 12              |
| 19           | 8                               | 3               |
| 20           | 12                              | 23              |
| 21           | 10                              | 10              |
| 22           | 20                              | 20              |
| 23           | 3                               | 16              |
| 24           | 6                               | 11              |
| 25           | 11                              | 4               |
| 26           | 4                               | 28              |
| 27           | 28                              | 6               |
| 28           | 1                               | 2               |
| 29           | 2                               | 26              |

Table 5  Comparison of the proposed method with individual feature selection algorithms used, in terms of the number of relevant features recommended in top 11 recommendations (K = 11) for ccRCC.

| Metric                        | Proposed Method | Random Forest | SFS | SBS | MIFS | Chi-Squared | F-score | JMI | NMI |
|-------------------------------|-----------------|---------------|-----|-----|------|-------------|---------|-----|-----|
| Number of relevant features   | 8               | 5             | 5   | 5   | 3    | 4           | 4       | 6   | 6   |
| recommended                   |                 |               |     |     |      |             |         |     |     |
| Percentage of relevant features recommended | 100%           | 62.5%         | 62.5% | 62.5% | 37.5% | 50%         | 50%     | 75% | 75% |

The values of the performance metrics are averaged for the values of K = 3,5,7,9,11 in Table 6. The individual results for K = 3,5,7,9,11 are shown in the form of a graph in Figure 3. RR@K is not shown separately in the graph, as it remains the same for all K. It can be seen from Table 6 and Figure 3 that the proposed method is superior to all the works mentioned above with higher average precision, precision, recall, F1, and equal or better RR.
Fig. 3 Performance comparison of the proposed method with SOTA methods on ccRCC dataset

| Method                | Average Precision@K | Precision@K | Recall@K | F1@K  | RR@K |
|-----------------------|----------------------|-------------|----------|-------|------|
| Baczkiewicz et al.[17]| 0.522                | 0.616       | 0.525    | 0.567 | 1.00 |
| Cataltepe et al.[18]  | 0.202                | 0.471       | 0.450    | 0.460 | 0.33 |
| Kanimozhi et al.[19]  | 0.357                | 0.606       | 0.525    | 0.563 | 0.50 |
| Kuanr et al.[20]      | 0.534                | 0.635       | 0.550    | 0.589 | 1.00 |
| Proposed method       | **0.663**            | **0.744**   | **0.650** | **0.694** | 1.00 |

### 3.5 Performance on Benchmark Datasets

To show the robustness of the proposed method, the performance of the proposed method is validated on four standard datasets: clear cell renal cell carcinoma(ccRCC), chronic kidney disease [29], Indian liver patient [30], and cervical cancer risk factors [31]. The results obtained are shown in Table 7.

The proposed method is compared with the next-best method, Kuanr et al. [20], (as in Table 6) for the benchmark datasets. The results are shown in Table 8. The proposed method out-performs Kuanr et al.[20] for all the four benchmark datasets.
Table 7  Overall performance of the proposed method on benchmark datasets. Average precision@K, precision@K, and recall@K are averaged for the values of K = 3,5,7,9,11. F1@K is calculated from precision@K and recall@K.

| Dataset                                | Average Precision@K | Precision@K | Recall@K | F1@K | RR@K |
|----------------------------------------|---------------------|-------------|----------|------|------|
| Cervical Cancer Risk Factors            | 0.892               | 0.913       | 0.443    | 0.597 | 1.00 |
| Chronic Kidney Disease                  | 0.669               | 0.755       | 0.318    | 0.448 | 1.00 |
| Clear Cell Renal Cell Carcinoma         | 0.663               | 0.744       | 0.650    | 0.694 | 1.00 |
| Indian Liver Patient Dataset            | 0.864               | 0.872       | 0.725    | 0.792 | 1.00 |

Table 8  Comparison of the proposed method with Kuanr et al., over four datasets. Precision@K and recall@K are averaged for the values of K = 3,5,7,9,11. F1@K is calculated from precision@K and recall@K.

| Dataset                                | Proposed Method      | Kuanr et al. [20]   |
|----------------------------------------|----------------------|---------------------|
|                                        | Precision@K Recall@K F1@K RR@K | Precision@K Recall@K F1@K RR@K |
| Cervical Cancer Risk Factors            | 0.913 0.443 0.597 1   | 0.873 0.429 0.575 1   |
| Chronic Kidney Disease                  | 0.755 0.318 0.448 1   | 0.646 0.271 0.383 1   |
| Clear Cell Renal Cell Carcinoma         | 0.744 0.650 0.694 1   | 0.635 0.550 0.589 1   |
| Indian Liver Patient Dataset            | 0.872 0.725 0.792 1   | 0.872 0.400 0.548 0.5 |

3.6 Statistical Testing

The proposed method has been statistically tested against the next-best method in Table 6. Table 8 shows the values of the performance metric for the proposed method and Kuanr et al. [20] for the four standard datasets. Wilcoxon signed-rank test, as a non-parametric test, has been used for studying the statistical difference between the two methods. This test has been used over the paired T-test as it does not require a normal distribution of the data. More details about the Wilcoxon signed-rank test can be found at [40].

The Z-score obtained for the paired observations in Table 8 is -3.02 which is lower than the value of the Z-score for a two-tailed hypothesis with a significance level of 0.05 (-1.96). Thus the difference between the two methods is statistically significant.

3.7 Classification Results

As an illustration, the ranking of features obtained from the eight feature selection methods is used to form a recommendation ensemble for the ccRCC dataset. This ensemble creates a positional feature recommendation table shown in column 3 of Table 4.

Once the final ranking of features is obtained, the two classifiers, NN and SVM, are used to investigate the discriminating power of the selected set of features. A standard NN classifier with 2-hidden layers (14 and 8 nodes) is used to assess how classification accuracy varies with each additional feature. Adam optimizer is used with a learning rate of 0.2. The hyper-parameters are chosen using a grid search methodology. For SVM, the one-vs-all multiclass classification technique is used with the RBF kernel.

The model developed using the features from the proposed feature recommender, for the ccRCC dataset, has also been compared with an existing work by Park et al.[7] that uses clinical features for renal cancer stage classification. In [7], 15 features were identified to be the most important ones, and they were used to do cancer stage
classification. Classification accuracy, for 10-fold crossvalidation, obtained using SVM and NN through [7] and those obtained using our proposed feature selection method are put in Table 9. From the table, it is seen that the earlier study yielded an accuracy of 82.3% and 76.3% using the SVM and NN classifiers, respectively. Whereas, our proposed feature recommendation method gives an accuracy of 96.6% and 98.6%, respectively with only 4 top-ranked features. This shows the superiority of the model built with features recommended through the proposed approach as compared to that of [7].

| Metric | Park et al.[7] | Proposed Method |
|--------|----------------|-----------------|
|        | SVM | NN | SVM | NN |
| Accuracy | 0.823 | 0.763 | 0.966 | 0.986 |

4 Conclusions

Proper stage identification or staging is a very important aspect of the treatment process of any disease. Diagnosis of a disease at an early stage can lead to better treatment options. A mutual Information and ensemble based automated feature selection and ranking strategy is proposed in this article. The performance of the proposed method is compared with four state-of-the-art methods and on four different datasets depicting superiority. The proposed method can identify and recommend risk factors that have the most discriminating power for disease diagnosis/staging. The risk factors recommended for ccRCC by the proposed method can be validated using an existing medical system (AJCC TNM) as well. The results obtained using the proposed method were shown to be statistically significant with respect to the next best method considered for comparison. This work can be extended to study medical images, extracting features, and recommending regions of interest for ease of analysis and diagnosis.

5 Acknowledgments

The authors are grateful to the Indo-US Science and Technology Forum (IUSSTF) for establishing a Virtual Networked Center and supporting a part of this research by sanctioning a project titled "Center for Distributed Deep Learning Framework for Classification" (File No IUSSTF/ JC-024/ 2018 dtd. 13.09.2019). A part of this work has been supported by the IDEAS - Foundation, The TIH at the ISI, Kolkata through sanctioning a Project No. ISI/TIH/2022/55/ dtd. Sept 13, 2022.
6 Declarations

6.1 Competing Interests
The authors declare that they have no conflicts of interest to report regarding the present study.

6.2 Availability of Data and Materials
Data used for this study is available from the corresponding author on request.

6.3 Author Contributions
Abhishek Dey, Debayan Goswami, Rahul Roy, and Susmita Ghosh designed the study. Abhishek Dey, Debayan Goswami and Rahul Roy collected the data, conducted the experiments and drafted the manuscript. Results were analyzed by all the authors. Susmita Ghosh, Yu Shrike Zhang, and Jonathan H. Chan enhanced and finalized the manuscript.

References
[1] Ricci F, Rokach L, Shapira B. In: Recommender Systems Handbook. vol. 1-35; 2010. p. 1–35.
[2] Alhassan AM, Wan Zainon WMN. Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis. IEEE Access. 2021;9:87310–87317. https://doi.org/10.1109/ACCESS.2021.3088613.
[3] Rado O, Ali N, Sani HM, Idris A, Neagu D. Performance Analysis of Feature Selection Methods for Classification of Healthcare Datasets. In: Intelligent Computing. Cham: Springer International Publishing; 2019. p. 929–938.
[4] Ceylan Z, Atalan A. Estimation of healthcare expenditure per capita of Turkey using artificial intelligence techniques with genetic algorithm-based feature selection. Journal of Forecasting. 2021;40(2):279–290. https://doi.org/10.1002/for.2747.
[5] Chen Y, Wang Y, Cao L, Jin Q. An Effective Feature Selection Scheme for Healthcare Data Classification Using Binary Particle Swarm Optimization. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME); 2018. p. 703–707.
[6] Halder A, Ghosh A, Ghosh S. Aggregation pheromone density based pattern classification. Fundam Inf. 2009 dec;92(4):345–362.
[7] Park K, Ryu K, Ryu K. Determining minimum feature number of classification on clear cell renal cell carcinoma clinical dataset. In: 2016 International Conference on Machine Learning and Cybernetics (ICMLC). vol. 2; 2016. p. 894–898.
[8] Bhavan A, Aggarwal S. Stacked generalization with wrapper-based feature selection for human activity recognition. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE; 2018. p. 1064–1068.

[9] Sahoo KK, Ghosh R, Mallik S, Roy A, Singh PK, Zhao Z. Wrapper-based deep feature optimization for activity recognition in the wearable sensor networks of healthcare systems. Scientific Reports. 2023;13(1):965.

[10] Sahoo AK, Pradhan C, Barik RK, Dubey H. DeepReco: deep learning based health recommender system using collaborative filtering. Computation. 2019;7(2):25.

[11] Waqar M, Majeed N, Dawood H, Daud A, Aljohani NR. An adaptive doctor-recommender system. Behaviour & Information Technology. 2019;38(9):959–973.

[12] Guo L, Jin B, Yao C, Yang H, Huang D, Wang F, et al. Which doctor to trust: a recommender system for identifying the right doctors. Journal of Medical Internet Research. 2016;18(7):e6015.

[13] Granda Morales LF, Valdiviezo-Diaz P, Reátegui R, Barba-Guaman L. Drug recommendation system for diabetes using a collaborative filtering and clustering approach: development and performance evaluation. Journal of Medical Internet Research. 2022;24(7):e37233.

[14] Chen J, Li K, Rong H, Bilal K, Yang N, Li K. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Information Sciences. 2018;435:124–149.

[15] Ge M, Ricci F, Massimo D. Health-aware food recommender system. In: Proceedings of the 9th ACM Conference on Recommender Systems; 2015. p. 333–334.

[16] Sayeb Y, Jebri M, Ghezala HB. A graph based recommender system for managing COVID-19 Crisis. Procedia Computer Science. 2022;196:348–355.

[17] Bączkiewicz A, Kizielewicz B, Shekhovtsov A, Wątróbski J, Salabun W. Methodical Aspects of MCDM Based E-Commerce Recommender System. Journal of Theoretical and Applied Electronic Commerce Research. 2021;16(6):2192–2229. https://doi.org/10.3390/jtaer16060122.

[18] Çataltepe Z, Uluayğmur M, Tayfur E. Feature selection for movie recommendation. Turkish Journal of Electrical Engineering & Computer Sciences. 2016;24:833–848. https://doi.org/10.3906/elk-1303-189.

[19] Kanimozhi G, Shanmugavadivu P, Rani MMS. Machine Learning-Based Recommender System for Breast Cancer Prognosis. In: Recommender System with Machine Learning and Artificial Intelligence. 1st ed. India: Wiley; 2020. p.
[20] Kuanr M, Mohapatra P. Outranking Relations based Multi-criteria Recommender System for Analysis of Health Risk using Multi-objective Feature Selection Approach. Data & Knowledge Engineering. 2023;145:102144. https://doi.org/10.1016/j.datak.2023.102144.

[21] Parmezan ARS, Lee HD, Spolaôr N, Wu FC. Automatic recommendation of feature selection algorithms based on dataset characteristics. Expert Systems with Applications. 2021;185:115589. https://doi.org/10.1016/j.eswa.2021.115589.

[22] Datta A, Ghosh S, Ghosh A. Wrapper based feature selection in hyperspectral image data using self-adaptive differential evolution. In: 2011 International Conference on Image Information Processing; 2011. p. 1–6.

[23] Banit I, Ouagua NA, Ait Kerroum M, Hammouch A, Aboutajdine D. Band selection by mutual information for hyper-spectral image classification. International Journal of Advanced Intelligence Paradigms. 2016 01;8:98. https://doi.org/10.1504/IJAIP.2016.074791.

[24] Elmaizi A, Sarhrouni E, Hammouch A, Nacir C. A new band selection approach based on information theory and support vector machine for hyperspectral images reduction and classification. In: 2017 International Symposium on Networks, Computers and Communications (ISNCC); 2017. p. 1–6.

[25] Nhaila H, Elmaizi A, Sarhrouni E, Hammouch A. A novel filter approach for band selection and classification of hyperspectral remotely sensed images using normalized mutual information and support vector machines. In: Proceedings of EMENA-ISTI 2018; 2019. p. 521–530.

[26] Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Computers in Biology and Medicine. 2019;112:103375. https://doi.org/10.1016/j.compbiomed.2019.103375.

[27] Gharsalli S, Emile B, Laurent H, Desquesnes X. Feature Selection for Emotion Recognition based on Random Forest. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016). vol. 4; 2016. p. 610–617.

[28] Novovicova J, Malik A. Information-theoretic feature selection algorithms for text classification. In: Proceedings of the International Joint Conference on Neural Networks. vol. 5; 2005. p. 3272–3277.

[29] Rubini L, Soundarapandian P, Eswaran P: Chronic kidney disease.

[30] Ramana B, Venkateswarlu N: ILPD (Indian liver patient dataset).
[31] Fernandes K, Cardoso J, Fernandes J.: Cervical cancer (risk factors).

[32] Akin O, Elnajjar P, Heller M, Jarosz R, Erickson BJ, Kirk S, et al. Radiology Data from The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma [TCGA-KIRC] collection. The Cancer Imaging Archive. 2016;https://doi.org/10.7937/K9/TCIA.2016.V6PBVTDR.

[33] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging. 2013;26/6:1045–1057.

[34] Silveira T, Zhang M, Lin X, Liu Y, Ma S. How good your recommender system is? A survey on evaluations in recommendation. International Journal of Machine Learning and Cybernetics. 2019;10:813–831.

[35] Swami U, Nussenzveig R, Haaland B, Agarwal N. Revisiting AJCC TNM staging for renal cell carcinoma: Quest for improvement. Annals of Translational Medicine. 2019;03:7:S18–S18. https://doi.org/10.21037/atm.2019.01.50.

[36] Laber DA. Risk factors, classification, and staging of renal cell cancer. Medical Oncology. 2006;23:443–454.

[37] Patel HD, Gupta M, Joice GA, Srivastava A, Alam R, Allaf ME, et al. Clinical stage migration and survival for renal cell carcinoma in the United States. European Urology Oncology. 2019;2(4):343–348.

[38] Ray RP, Mahapatra RS, Khullar S, Pal DK, Kundu AK. Clinical characteristics of renal cell carcinoma: Five years review from a tertiary hospital in Eastern India. Indian Journal of Cancer. 2016;53(1):114–117.

[39] Warren AY, Harrison D. WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: standards and controversies. World Journal of Urology. 2018;36:1913–1926.

[40] Woolson RF. Wilcoxon signed-rank test. Wiley Encyclopedia of Clinical Trials. 2007;p. 1–3.