A hybrid filter-based genes selection stochastic model for cancer prediction and classification based on gene expression data

WAHID, Abdul, HABIB, Sadaf, ALAM, Urooj, HASSAN, Waqar and AKMAL, Muhammad <http://orcid.org/0000-0002-3498-4146>

Discover Artificial Intelligence

RESEARCH Open Access

A hybrid filter-based genes selection stochastic model for cancer prediction and classification based on gene expression data

Abdul Wahid[1], Sadaf Habib[2], Urooj Alam[3], Waqar Hassan[4] and Muhammad Akmal[5*]

*Correspondence:
Muhammad Akmal
m.akmal@shu.ac.uk
[1]Department of Mathematics, Air University Islamabad, Islamabad 44000, Pakistan
[2]Department of Mathematics and Statistics, University of Southern Punjab, Multan, Pakistan
[3]Medside Healthcare, Sandy Spring, Georgia 30350, USA
[4]Protection and Control Department, Eversource Energy, 247 Station Drive, Westwood, USA
[5]School of Engineering & Built Environments, Sheffield Hallam University, Sheffield S1 1WB, UK

Abstract

The study of high-dimensional genes expression data has made an important role in disease diagnosis and cancer type classification. Nevertheless, due to the problem of curse of dimensionality, dimension reduction methods, particularly feature selection (FS) methods, are crucial for eliminating redundant genes and improving disease classification. Stochastic FS models are vital class of FS models in the analysis of genes expression data. Despite the availability of some stochastic FS models, each model has its own limitations. In this study, we propose a hybrid stochastic FS model that embedding multivariate filtering in a hidden Markov-model (HMM) framework. It addresses the problem of redundancy and applicable for binary and multi-class problems, therefore, we named Generalized Multivariate FS based HMM (GMFS-HMM). On colon data, the new GMFS-HMM model outperforms the state-of-the-art HMM model, achieving the highest accuracies (>90%) for Randon Forest (RF) and across 10 to 30 selected genes. Gene enrichment analysis of colon data further validated this performance of GMFS-HMM. Moreover, the applicability and performance of the GMFS-HMM was also demonstrated on three multi-class datasets, including high-dimensional RNA-seq Pan-cancer data. The analysis on benchmark datasets illustrates that GMFS-HMM has improved classification accuracy and applicability on binary as well as multi-class datasets.

Keywords Alzheimer's disease, Hidden Markov-model, Feature selection, Redundancy, Genes expression data, Classification, Cancer prediction

1 Introduction

In past two decades, the innovations in microarray high-throughput technology have empowered collection of massive amounts of genes encoded genetic information. The abnormalities and variations in these genes have a major role in human diseases such as cancer, Alzheimer and neuro degenerative disorders. For a specific gene, the variations in expression levels can be measured and used as genetic traits for disease diagnosis. Many of genes are irrelevant and redundant to human disease and make no importance to diagnosis. Therefore, selecting subset of highly discriminative features or genes are very crucial in constructing machine learning models for disease diagnosis and drug

development [1]. However, the problems of high-dimensionality [2] of microarray data with small sample size deteriorate the performance and increase complexity of disease classification models [3].

Since the main challenges with large-scale microarray data is the existence of redundant and non-informative genes that do not contribute to disease status and need to be discarded. To deal with these challenges of microarray data, dimensionality reduction techniques [4] are widely used to remove non-informative genes and improve disease classification. Techniques for dimension reduction fall into two categories: feature extraction [5, 6] and feature selection(FS) [7–9]. FS finds a small subset of the original features or called genes, while feature extraction uses transformation technique to obtain reduced space from the original features or contained all information of original features. In many studies, using FS methods has been found to provide better performance compared with feature extraction methods for addressing the challenges of high-dimensional microarray data [10, 11].

FS approaches are further categorized as filter, wrapper, embedded and hybrid models. Filter models independently choose genes or features from the learning classifier, while the rest use learning algorithms to assess a subset of chosen features [12]. Filter models are computationally fast but produce low classification accuracy, wrappers are computationally expensive but produce higher accuracy than filter models, and embedded models perform better than aforementioned models in terms of accuracy but have high computational cost than filter models. Currently, hybrid models have emerged to unify two FS approaches from the same or different categories, aiming to use advantages of diverse models. In this article, we focus on filter models with stochastic properties.

In microarray data analysis, most of the existing filter models [13] select subset of genes without considering correlation or genetic redundancy among genes. It refers to the situation where large number of genes performs the same function, which results non-predictive models with low efficiency, high computational cost and biased findings. Such filter methodologies often rank individual genes according to their separate importance score and ignore the degree of correlation among them. A crucial problem in high-dimensional gene expression data mining is to allow medical geneticist with a predictive filter models that significantly determines and selects non-redundant genes. Another problem for geneticist to develop novel filter models that feasible for both binary and multi-class genes expression data. Multi-class microarray data classification is a hard problem than binary classification[14]. Recently, in [15] the authors propose a filter FS approach by employing hidden Markov model based on five univariate feature ranking methods. The major limitation of the last approach is to avoid feature redundancy, non-linearity in data, and infeasibility for multi-class classification problems.

In the increasing use of integrated hybrid models, constructing from various FS models to develop more robust and predictive models has gained attention recently [64–66]. However, a single FS model has not performed well for complex and large-scale data because an individual FS model is insufficient, has high instability, and has low accuracy [65]. Thus, key advantages of integrating various FS models have the ability to deal with challenges of data obtained from different sources, for instance, big data from multi-modalities [67]. Motivated by such modelling, a model is being developed in this article to capture non-linearity, redundancy, and multi-class labels problems. Unlike traditional techniques, the new model is integrated by well-known FS methods through stochastic

processes, instead of simply combining their final individual results, for example, in ensemble learning [68, 69]. These models in the unified FS proposed approach have unique characteristics like simplicity, high accuracy, and low computational costs. Different from existing integrated models, another advantage is to compare the similarity of among different FS models in terms of evaluation metrics. This intuition unfolds new directions in the future. The main contributions of the present work can be summarized as follows:

- We adopt a stochastic FS method based on information theory designed for binary as well as multi-class classification problems.
- The new approach is designed to plug-in one multivariate (mRMR) and two univariate (IG and GI) filter models into the layout of HMM to address redundancy and capture non-linear relationships between a response and features.

The remaining article is divided in to four sections. In section 2, we will review some recent hybrid FS methods that have been published in the literature. The new model for FS is outlined in section 3. Section 4 presents the application of novel hybrid model to four real-world genes expression datasets. Finally, section 5 describes conclusions and future work.

2 Related literature

According to the strategy of FS, the machine learning models in supervised learning are allocated into four categories: filter, wrapper, embedded and hybrid [4]. In this section, we overview on recent supervised filter and hybrid approaches and their applications in genomics data.

2.1 Filter FS methods

Numerous filter FS models have been developed to tackle the challenges of high-dimensional microarray data. Due to the simplicity and computational efficiency in addressing microarray data challenges, filter models have attracted interest for handling gene selection issues [16, 17]. At early stage, [18] proposed a filter FS model based on correlation analysis among features, but it has not been efficient for very large-scale data sets. Many univariate filter models have been proposed to select relevant features [19–21]. Although easy to implement, these models have been observable drawback concerning redundancy among features. To avoid redundancy, many state-of-the-art multivariate filter FS models [22–24] have been proposed for microarray datasets. [25] designed biomarker identifier (BMI), a filter genes selection method to select markers for high-dimensional lung cancer data. They unified three statistical approaches, i.e., distribution theory, variance and logistic regression, into a joint framework that best distinguish between samples with and without lung cancer. In order to cope with the high false discovery rate and redundancy problems, [26] developed a filter model for paired microarray gene expressions analysis. The authors have shown that their proposed model performs better than compared methods in terms of accuracy and stability with gene expression data. [27] introduced a filter model based on Chebyshev distance-outlier identification, an improved version of Relief-based algorithms [28] that deals with uncertainty and noises. None of these methods consider the significance of every feature and category within the data. To overcome this problem, [29] was proposed a pre-filter algorithm by

using features weights obtained from decision tree to improve classification accuracy. Recently, the authors in [30, 31] presented new filter FS models and demonstrated their applications in genomics. Moreover, some other relevant works can be found in [9].

2.2 Hybrid FS methods

Here, we review supervised hybrid FS methods based on multiple filter approaches. Initially, [32] suggested a hybrid FS technique that integrates filter algorithms for addressing classification problems. Another filter methods-based hybrid FS approach, namely, MFHFS was developed in [33]. This methodology consists of five stages, in first stage; the samples are normalized and discretized by applying equal width interval binning algorithm [34] with 10-fold cross-validation for discarding outliers and noises. The second is variables combination stage, where multiple filters were used for obtaining best subsets and determine two types of weights [33]. In stage 3, called, feature refinement stage in which redundant information is filter out to find final optimal subset. In last stage, Q-range approach is implemented to improve computation. This approach outperforms the traditional filters and hybrid techniques. Recently, in [35], a two-stage hybrid FS technique was introduced for high-dimensional time series data. First, three filter models were unifying for efficient feature selection and secondly, the Levy's flight [36] was implemented for lowering computational complexity. Furthermore, two more related works based on this same concept were developed in articles [37, 38].

In recent studies, HMM [15] is one of the hybrid FS models in which dimension of data is reduced based on combining five ranking methods into HMM framework. Besides, some good characteristics of HMM, it had several disadvantages due to the use of five weak features ranking techniques. In this study, the problems of HMM method are addressed and developed new methodology with better feasibility to high-dimensional genes expression data.

3 Proposed feature selection method

In this section, an optimum feature selection methods based HMM filter approach was presented to address the problems of [15] for gene microarray datasets. The main disadvantages of [15] are that it is infeasible for multi-class classification problems and adopted for univariate feature selection methods that unable to capture redundancy among genes. To overcome these problems and select disease relevant genes, we propose a hybrid model that considers redundancy among genes and feasible in binary as well as multi-class data. In this approach, the following three multivariate feature selection methods are built in HMM as hidden states.

3.1 Information gain

This method [39] uses information-theoretic idea of entropy for evaluating features and commonly used in the area of machine learning, text and cancer classification. It measure the association between two features and identify most relevant features in filter scheme and computationally very fast. In [39], the entropy of a feature X is computed as follows:

$$H(X) = -\sum_{x} p(x)log(p(x)). \tag{1}$$

The information contained in X subsequently observing the results of other feature Y, called conditional entropy, is given by

$$H(Y/X) = -\sum_{y}\sum_{x} p(x,y)log(p(x/y)) \tag{2}$$

where $p(x)$ denotes the prior probability of X and $p(x/y)$ denotes the posterior probability of X given Y. Using Eq.(1) and Eq.(2), the information gain (IG) is formulated as,

$$IG(X,Y) = H(X) - H(Y/X). \tag{3}$$

IG provides the degree of association between two features, when two features are statistically independent then IG will be zero otherwise greater than zero. Hence, larger value of IG indicates high redundancy or relevancy between two features. Fortunately, IG can be applied for both continuous and discrete features.

3.2 Minimum redundancy-maximum relevance

The minimum redundancy-maximum relevance (mRMR) [40] is a powerful filter feature selection algorithm based on information-theoretic criterion. By using this method, top sets of features are selected and represent relevance of response variable efficiently. This method maximizes the relevance of features with response variable and minimizes redundancy features. To find the set of features $\mathcal{S}$ with x_i on decided class c of output variable. The relevance scores for each feature and output variable can be calculated by maximizing dependency as follows:

$$max\mathcal{M}(\mathcal{S};c), \quad \text{where} \quad \mathcal{M} = \frac{1}{\mathcal{S}}\sum_{x_i\in\mathcal{S}} IG(x_i;c) \tag{4}$$

On the other hand, to reduce the redundancy problem among features, the features set $\mathcal{R}$ with minimum redundancy can be computed as:

$$min\mathcal{R}(\mathcal{S}), \quad \text{where} \quad \mathcal{R} = \frac{1}{\mathcal{S}^2}\sum_{x_i,x_j\in\mathcal{S}} IG(x_i;x_j) \tag{5}$$

Furthermore, Eq.(4) and Eq.(5) are combined to obtain the mRMR method, as

$$max\{\psi(\mathcal{M};\mathcal{R})\}, \quad \text{where} \quad \psi = \mathcal{M} - \mathcal{R}. \tag{6}$$

It has range [0, 1].

3.3 Gini index

Gini index (GI) [41] is used to measure heterogeneity and impurity and efficient for high-dimensional data processing and classification. If a feature is randomly labeled according to some distribution in a data, GI can identify it as incorrectly classify. Many authors have used GI for the goal of feature selection [60]. However, Gini index can be calculated by using the following formula:

$$GI(\mathcal{N}) = 1 - \sum_{i=1}^{C} p_i^2 \tag{7}$$

where $\mathcal{N}$ represents the collection of data samples which defines different classes (i=1, 2,$\cdots$, C) and p_i is the probability of any sample in i= 1, 2, $\cdots$, C. GI ranges from 0 to 1, 0 represents homogeneity and 1 shows impurity, means samples are allocate among different classes.

3.4 HMM-Hidden Markov model

HMM was introduced by [43] in the late 1960's. There are many applications of HMMs in real-word, for example, initially it was commonly applied in speech recognition. Afterwards, HMM were used in analysis of biological sequences since late 1980's [44]. HMM is a sequential statistical model used to know how hidden information from observed variables is? HMMs are probabilistic models that are sequence of random variables and use Markov process to find unknown parameters.

HMM can be computed by its five components including hidden states, observed states, emission probabilities, transition probabilities, and initial probability. The description of HMM model in the context of FS includes hidden states which show the probability of chosen features by every method. Observed sequence represents the selected features position by each FS method while transition matrix consists of probabilities of any feature hidden state effect the next feature state. The emission matrix shows the probability that any observed feature is produced in a state depends on the features ranking of a FS method. Finally, the probabilities of starting states (t=0) are denoted the initial state probability distribution. In the context of proposed methodology, these components are defined as follows:

- Hidden states are denoted as, $S_t = \{S_1, \cdots, S_N\}$. Which represents the set of all features that rank with each model.
- Observations(or features) of the HMM model is denoted as, $X_t = \{X_1, \cdots, X_T\}$.
- The transition state probability distribution, also known as probability transition matrix $A = \{p_{ij}\}$, describe the probabilities of transitioning from state S_{t-1} to S_t. It can be explained as:

$$p_{ij} = Pr\{S_{t,j} = 1 | S_{t-1,i} = 1\}$$

where S_t denotes the current state and each row sums equal to 1, i.e., $\sum_{j=1} p_{ij} = 1$.

- The emission state probability distribution also called emission probabilities is (T$\times$ N) matrix, its elements B_j demonstrate the probabilities of making observation $X_{t,n}$ given $S_{t,j}$, i.e.,

$$B_j^n = Pr\{X_t = n | S_t = j\}, \ \ 1 \leq j \leq N, \ \text{ and } \ 1 \leq n \leq T.$$

Where *n* is *n*th observation and X_t is observed state at time t. For B_j^n, we must have

$$B_j^n \geq 0, \ \ 1 \leq j \leq N, \ \text{ and } \ 1 \leq n \leq \xi$$

$$\sum_{j=1}^{N} B_j^n = 1, \ \ 1 \leq j \leq N.$$

- Initial state distribution is N$\times$1 vector of probabilities which is given by

$$\pi_i = Pr\{S_{1i} = 1\} \text{ with } Pr\{S_1|\pi\} = \prod_{i=1}^{N} \pi_i$$

These five parameters (N, ξ, B, p_{ij}, π) completely explained the HMM. In proposed method, hidden states shows that the relevant features are associated to some specified model (FS method) and sequence of observation is obtained by hidden states which gives feature rank resulted from selected subsets. Figure 1 shows the relationship between hidden states and observed states (or observations).

3.4.1 Transition matrix

Transition matrix, denoted by A, gives the probabilities between hidden states that are equal to the common top-ranked features, obtained by applying different FS models on training dataset. Top-ranked features are selected by using three different multivariate feature ranking method. Transition matrix is established by intersection of these top-ranked features. The numerator for method i and method j is defined as:

$$\mathcal{M}_i \cap \mathcal{M}_j$$

According to HMM property, the total of each row in transition matrix must be equal to 1. Therefore, each element of transition matrix A is transformed to normalized score by considering observations among various hidden states of HMM. All elements in A represents the pairwise overlapping genes or features produced by each FS method. Normalized scores are computed by dividing the number of overlapping genes between two different methods by the total number of overlapping genes between a given FS method and remaining all methods. The sum for a given ith method and common over all remaining methods ($j = 1, 2, \cdots$) is computed as follows:

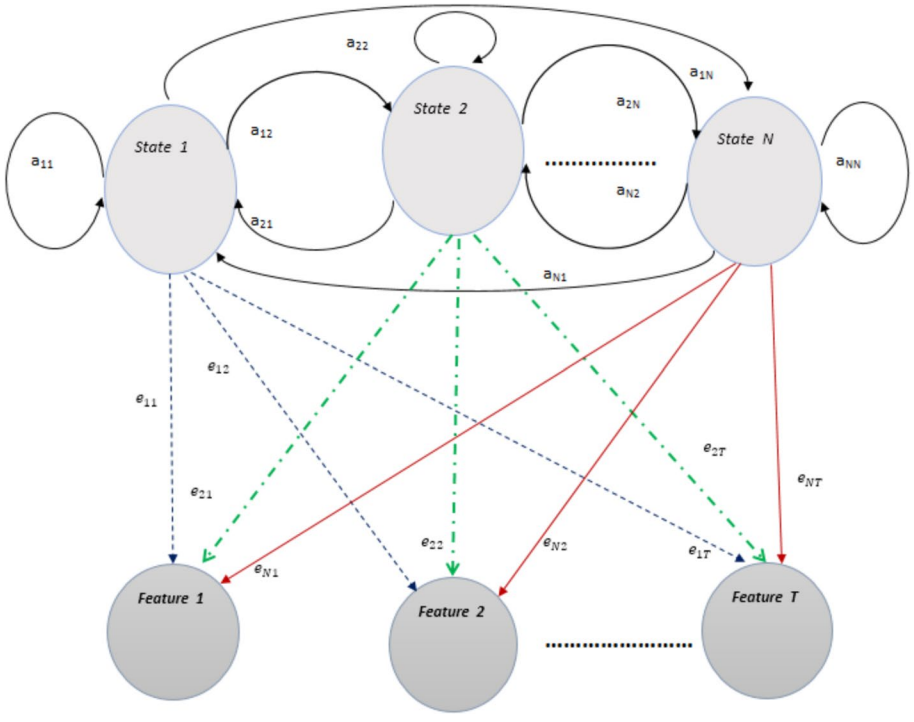

Fig. 1 Diagram of the Hidden Markov model

$$sum_i = \sum_{i \neq j} \mathcal{M}_i \cap \mathcal{M}_j$$

Hence, the transition matrix is obtained from last two equations. Specifically for our methodology, the sum of overlapping genes score for IG, mRMR and GI are calculated and further using these scores for computing transition matrix A as:

$$A = \begin{array}{c|ccc} & \text{S 1} & \text{S 2} & \text{S 3} \\ \hline \text{S 1} & 0 & \dfrac{(\mathcal{M}_1 \cap \mathcal{M}_2)}{sum_{IG}} & \dfrac{(\mathcal{M}_1 \cap \mathcal{M}_3)}{sum_{IG}} \\ \text{S 2} & \dfrac{(\mathcal{M}_2 \cap \mathcal{M}_1)}{sum_{mRMR}} & 0 & \dfrac{(\mathcal{M}_2 \cap \mathcal{M}_3)}{sum_{mRMR}} \\ \text{S 3} & \dfrac{(\mathcal{M}_3 \cap \mathcal{M}_1)}{sum_{GI}} & \dfrac{(\mathcal{M}_3 \cap \mathcal{M}_2)}{sum_{GI}} & 0 \end{array} \quad \text{for } i = 1, 2, 3.$$

By doing so, the values in the main diagonal of the transition matrix are 0 and every row sum becomes equal to 1.

3.4.2 Emission matrix

Probabilities in emission matrix are computed as the relationship between hidden states (FS methods) and the observations (genes or features). The probabilities in emission matrix shows that, according to the rank of features by each method, features are omitted in state *i*. In emission matrix, top-ranked features are indicated by the high score.

$$\phi^{-R(i,j)}$$

where ϕ is the tuning parameter which means it can be adjusted for different datasets. Value of ϕ must be appropriate, not more than 4 and below 1 [15]. If $\phi < 1$, then top-ranked feature has less probability of selection due to reduced score, on the other hand, if $\phi > 4$, then score of high-ranked feature is increased. The gap between score is expand, which causes misleading results. So, the optimum range of ϕ is between 1 and 4. For more detail, the analysis of ϕ between the 1 and 4 range, and its impact on the performance of proposed model GMFS-HMM on Colon data is discussed in Sect. 5.

Formation of emission matrix for *i* th method(state) and *j*th feature is given by:

$$E = [\xi_{i,j}] = \frac{\phi^{-R(i,j)}}{\sum_{k=1}^{p} \phi^{-R(i,k)}} \tag{8}$$

where $R(i, j)$ is the rank of *j*th feature for *i*th method and p indicates total number of features. Emission matrix with 3 methods in rows and p features in columns can be written as:

$$E = \begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2p} \\ \xi_{31} & \xi_{32} & \cdots & \xi_{3p} \end{pmatrix}$$

3.5 Computation algorithm

In this part, we use the well-known expectation-minimization (EM) algorithm, called Baum-Welch algorithm, to estimate the unknown parameters $\varphi = \{A, E, \pi\}$ of the proposed HMM filter FS model. Our objectives are to find optimum values for observations (or genes) and rank them according to every FS method in the training set. It updates the

estimation of initial, transition and emission probabilities (steps 2 and 3). The forward and backward variables are calculated and updated in step 4. Furthermore, in step 5, we calculate variables $\tau_{i,j}(t)$ and $\gamma_t(j,l)$ (posterior probability). The expectation-maximization procedure is employed to find optimum φ^* utilizing steps 4 and 5. The whole process is repeated until convergence. The pseudo codes of the Baum-Welch computation are given in the following Algorithm.

Baum-Welch Algorithm

Input: HMM model $\varphi = \{A, E, \pi\}$ and sequence of observations $X = x_1, \cdots, x_T$
Output: φ^* consisting of optimal values of A, E and π to maximize the probability $Pr(\phi/X)$
step 1: Initialization
$\alpha_i = \pi_i b_i(x_1)$, $B_T(i) = 1$, and $1 \le i \le N$
Step 2: Iterate
Step 3: Estimation step
Step 4: Forward-backward iterative computation

$$\alpha_{t+1}(i) = b_i(x_{t+1}) \sum_{i=1}^{N} \alpha_t(i) \cdot a_{ij} \text{ and}$$

$$B_t(i) = \sum_{i=1}^{N} B_{t+1}(i) a_{i,j} b_i(x_{t+1}). \text{ for } 1 \le i \le N \text{ and } 1 \le t \le T-1$$

Step 5: Computing $\gamma_t(i)$, $\gamma_t(j)$ and $\tau_t(i,j)$ where,

$$\tau_t(i,j) = \frac{\alpha_i(t) \cdot \alpha_{i,j} \cdot b_j(x_{t+1}) \cdot B_{t+1}(j)}{\sum_i^N \sum_j^N \alpha_i(t) \cdot \alpha_{i,j} \cdot b_j(x_{t+1}) \cdot B_{t+1}(j)}$$

$$\gamma_t(i) = \sum_{j}^{N} \tau_t(i,j) \text{ and,}$$

$$\gamma_t(j) = \frac{\alpha_t(j) \cdot B_t(j)}{\sum_j^N \alpha_t(j) \cdot B_t(j)}, \text{ for } 1 \le j \le N \text{ and } 1 \le t \le T-1$$

Step 6: Maximization
Step 7: Computing optimal parameters

$$\alpha_{i,j} = \frac{\sum_{t=1}^{T-1} \tau_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \text{ for } 1 \le i,j \le N,$$

$$b_{i,j} = \frac{\sum_{t=1}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \text{ and}$$

$$\pi_i = \gamma_1(i), \text{ for } 1 \le i \le N.$$

Step 8: set $\varphi \to \varphi^*$

3.6 Feature ranking

After obtaining the optimum parameters in φ by applying Baum-Welch iterative algorithm, we consider emission matrix *E*. Adding the values of emission matrix columnwise (we add the values for each hidden state or FS method), the score which is obtained, taken as feature importance score. Then feature or gene with the highest score will be selected first as most important one, similarly the second feature with second highest score, and so on. We also compute mean value of emission matrix for all states or methods (i = 1, 2, $\cdots$, n) and determine the scores for each feature. For *j*th feature, we compute score as follows:

$$score_j = \frac{1}{n} \sum_{i=1}^{n} \xi_{ij}, \text{ for } j = 1, 2, \cdots, p \tag{9}$$

where ξ_{ij} represents the elements of emission matrix for *i*th method and *j*th feature estimated by Baum-Welch algorithm. Finally, the scores of all features are computed in Eq.(9) and are ranked in descending order of magnitude and select a subset of top-ranked subset of features.

3.7 Example

As a demonstrating hypothetical example in Fig. 2, we take a set of four features $\{f_1, f_2, f_3, f_4\}$ shown in step 1. Calculate feature ranking matrix by applying three feature selection methods $\{M_1, M_2, M_3\}$. Transition matrix is obtained by intersection of features given in steps 3 and 4. The overlap score among three selection methods are 1, 1, and 2, respectively. We take 50% features out of total 4 features that is comprise the first two features. For example, the overlap score between two methods M_1 and M_2 is 1 of first two features (i.e., 50% features) which is one common feature that is f_4. Emission matrix is calculated according to the feature ranking matrix using three FS methods. For example, ξ_{12} in step 8 with tuning parameter ϕ=2 is computed by

$$\frac{\phi^{-R(1,2)}}{\sum_{k=1}^{p}\phi^{-R(1,j)}} = \frac{2^{-4}}{0.94} = 0.067$$

where $R(1,2) = 0.063$ and denominator is $2^{-R(1,1)} + 2^{-R(1,2)} + 2^{-R(1,3)} + 2^{-R(1,4)} = 0.94$, as shown in steps 6 and 7 of Fig. 2, respectively. High score of a feature indicate that the feature is more significant and important. In this example the feature selected at first place is f_3, in second place is f_4 and so on.

4 Numerical studies

In this section, we assess the performance of proposed FS method on four high-dimensional gene expression datasets.The classification accuracy can be affected after selecting a subset of genes by FS methods. A method will be performed better if it selects discriminative genes and discard the redundant ones. The classification accuracy and precision are computed on selected subsets of genes by three well-known classifiers, including Random Forest (RF)[45], Support Vector Machine (SVM), and k-Nearest Neighbor(kNN).These classifiers have been applied using R packages, **class**, **e1071**,

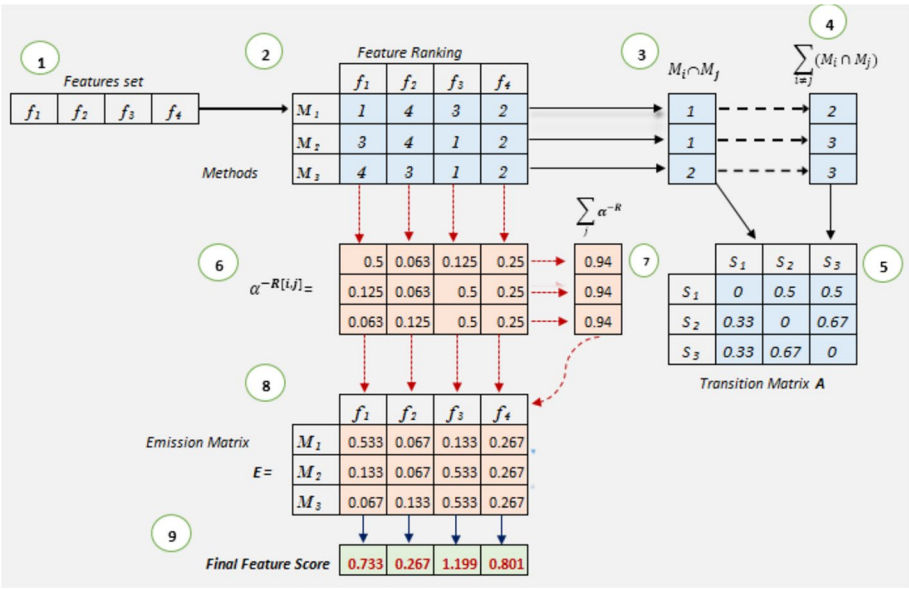

Fig. 2 An Empirical Illustration of the proposed methodology consisting of: 1. Original features set; 2. Feature ranking; 3. intersection of features; 4. Sum of intersection of features; 5. Transition matrix; 6. Probability for ranking to get emission matrix; 7. Adding probabilities; 8. Emission matrix; 9. Final feature importance score

randomForest for kNN, SVM and RF, respectively. Computing environment utilized Windows 10 OS PC, an Intel i7 CPU with 24 GB memory of Apple M4 Pro (Mac mini).

Before employing the FS methods to the datasets, some data wrangling need to be performed with parameter settings.

4.1 Data pre-processing

Data wrangling is necessary as it infer how to analyze large-scale gene expression data and prepare it for estimation, computation, information graphics and visualization. The genes expression datasets are primarily raw datasets and preprocessed by replacing missing and NA's values with average values, and eliminating mistype symbols before implementation of learning models. `Filter()` row function is used to filter out only that rows that are relevant to our criteria and exclude unexpected symbols, false and NA values from dataset. The R package **dplyr** is used for wrangling genes expression data used in this study. For scaling and control variation, the following Z-score normalization is used:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where μ and σ are the mean and variance of the ith feature or gene x_i.

Furthermore, for disease prediction, the samples were divided into two parts: training and testing sets. Let $D^{train} = (X^{train}, Y^{train})$ and $D^{test} = (X^{test}, Y^{test})$ denote the data in the training and test sets, respectively. Where X represents the features (or genes) matrix while Y represents class labels variable. The training dataset is used for fitting the models while test data is used for classification prediction. In this analysis, 80%:20% or 70%:30% training:testing ratios are used. For further validation, 5-fold and 10-fold cross-validation repeated over 50 times is used for three classifiers. The results of the folds are averaged and rounded up to 3-significant digits.

4.2 Evaluation metrics and parameters setting

Accuracy is widely used as an evaluation metric to estimate/measure the performance of feature selection methods for binary and multi-class classification problems. For evaluation, we used a confusion matrix to generate different evaluation metrics for a classifier after selecting a subset of features from filter learning models. In confusion matrix, there are four important expressions and their definitions are as follows:

Table 1 Accuracy±2SD results for colon dataset at different selected subset of genes

	kNN		SVM		RF	
Genes	***HMM***	***GMFS-HMM***	***HMM***	***GMFS-HMM***	***HMM***	***GMFS-HMM***
5	0.549±0.016	**0.6**±0.008*	0.589±0.019	**0.661**±0.015*	**0.619**±0.022	0.608±0.024
10	0.795±0.018	**0.845**±0.015*	**0.928**±0.091*	0.799±0.012	0.802±0.021	**0.939**±0.025*
15	0.851±0.065	**0.853**±0.075	**0.907**±0.020*	0.853±0.011	0.923±0.02	0.923±0.024
20	0.235±0.068	**0.573**±0.076*	0.179±0.023	**0.438**±0.012*	0.106±0.021	**0.442**±0.028*
25	0.816±0.060	**0.844**±0.039	**0.906**±0.018	0.8890.014	0.931±0.022	**0.935**±0.024*
30	0.866±0.063	**0.872**±0.038	**0.901**±0.019	0.897±0.020	0.912±0.011	**0.934**±0.006*
Max	0.866	**0.872**	**0.928**	0.897	0.931	**0.939**
Min	0.235	**0.573**	0.179	**0.438**	0.106	**0.442**
Mean	0.685	**0.765**	0.735	**0.756**	0.716	**0.797**

* Statistical significant difference of results at 0.05 level of significance

- True Positive (TP): there are instances in which the predicted **yes** actually included to the class **yes**.
- True Negative (TN): there are instances in which the predicted **No** actually included to the class **No**.
- False Positive (FP): there are instances in which the predicted **Yes** actually included to the class **No**.
- False Negative (FN): there are instances in which the predicted **No** actually included to the class **Yes**.

Accuracy of a classifier can be computed from the confusion matrix using formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision is also used to evaluate the worth of classification. It is the proportion of TP to the sum of TP and FP. It can be calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity or recall is the ratio of TP to the summation of TP and FN and can be computed as follows:

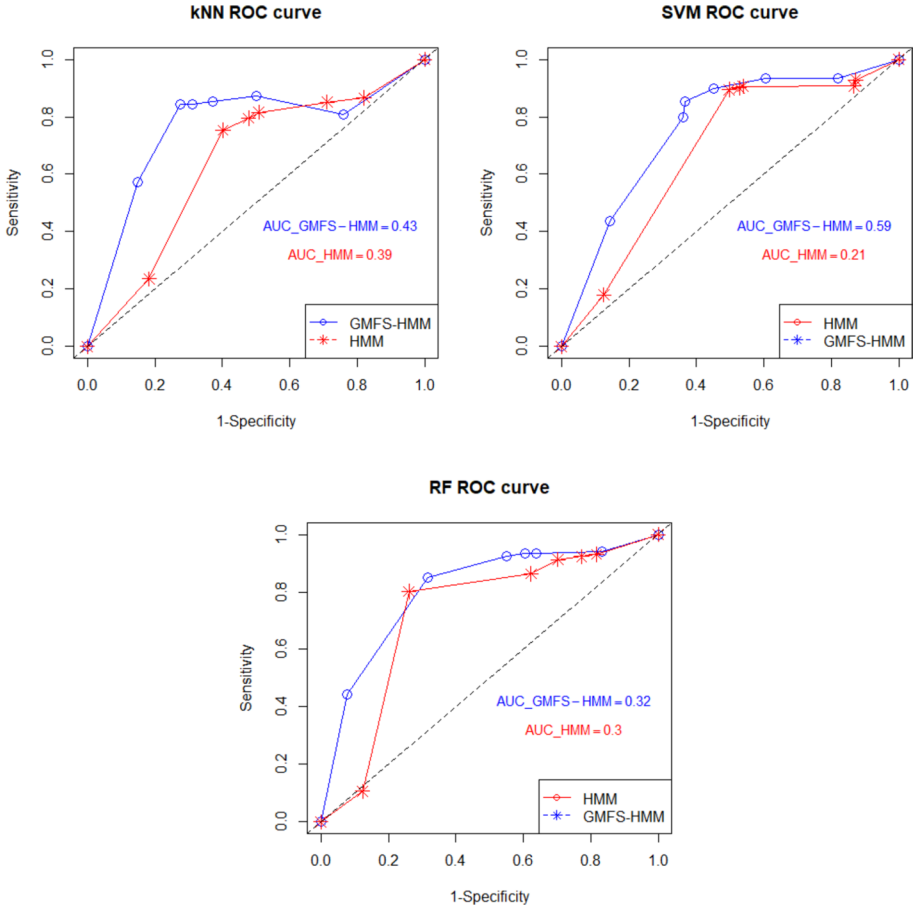

Fig. 3 Receiver operating characteristic curves with AUC values of two FS methods evaluated by three classifiers for colon data

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity is the ratio of TN divided by the sum of TN and FP. It can be calculated as:

$$Specificity = \frac{TN}{TN + FP}$$

Furthermore, the later two metrics have combined and find the Receiver Operating Characteristic (ROC) curve (in terms of sensitivity and specificity) to assess the performance of selection methods.

However, various parameters significantly influence the performance of FS models and classifiers. In this article, we set the number of selected genes as {5, 10, 15, 20, 25, 30} for the first two datasets while setting {10, 20, 30, 40, 50, 60, 70} for the last data.The Random Forest (RF) is performed with its two parameters, ntree and mtry. The parameter ntree is represented as number of trees and mtry used for the number of variables of random samples at each division or cut. We use ntree = 100, mtry = 1 and remaining are kept by default. For SVM, we used Radial Basis Function kernel with default parameters. In classifier k-NN, k is selected as $\sqrt{n}$, where n is the total number of samples or patients. Since model GMFS-HMM integrate three FS approaches and each state was given equal importance, the initial state probability distribution is assumed to be $\pi = (1/3, 1/3, 1/3)$. Finally, in the proposed model, the value of parameter ϕ is chosen as 2 (see section 5 for details).

4.3 Application to gene expression datasets

This section applies the proposed method to the four real-world biomedical datasets and comparing the results with the state-of-the-art method [15].

Example 1 **(Colon Dataset)**: The binary-class colon data [46] contains 2000 genes expression and 62 samples (tissues) with 40 colon tumor tissues and 22 normal tissues. This data is also available in R package **RaSEn**. Table 1 shows the classification performance of pre-specified subsets of genes selecting by two FS models using three classifiers.

Table 2 Gene enrichment analysis of top 15 selected genes for colon data

Rank	Gene ID	Gene Description
1	T54364	PROTEASOME COMPONENT C2 (HUMAN)
2	T71025	Human(human)
3	H85835	PROTEIN (CAENORHABDITIS ELEGANS)
4	M11220	Granulocyte-macrophage colony stimulating factor (GM-CSF mRNA)
5	H85528	RETINOBLASTOMA-LIKE PROTEIN 1 (Homo sapiens)
6	X66839	H.sapiensMaTu MN mRNA for p54/58N protein
7	H08393	COLLAGEN ALPHA 2(XI)CHAIN (Homo sapiens)
8	T88902	COT PROTO-ONCOGENE SERINE/THREONINE-PROTEIN KINASE (Homo sapiens)
9	X51416	Human mRNA for steroid hormone receptor hERR1
10	H43887	COMPLEMENT FACTOR D PRECUSOR (Homo sapiens)
11	U30872	Human mitosin mRNA, complete cds
12	L20688	Human GDP-dissociation inhibitor protein (Ly-GDI) Mrna, complete cds
13	M63391	Human desmin gene, complete cds
14	X57346	H.sapiens mRNA for HS1 protein
15	L05144	PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC(HUMAN);contains Alu repetitive element; contains elements PTR5 repetitive element

Table 3 Results comparison of three classifiers on RNA-seq data in terms of average accuracy

	No.of Selected Features by GMFS-HMM			
Classifiers	**50**	**100**	**150**	**200**
kNN	0.9835±0.009	0.9962±0.004	0.9946±0.006	0.9960±0.005
SVM	0.9825±0.010	0.9931±0.006	0.9918±0.007	0.9929±0.006
RF	0.9775±0.011	0.9848±0.009	0.9814±0.010	0.9841±0.010

Table 4 Average accuracy results of three classifiers on RNA-seq data after using SMOTE

	No.of Selected Features by GMFS-HMM			
Classifiers	**50**	**100**	**150**	**200**
kNN	0.9832±0.007	0.9979±0.002	0.9968±0.003	0.9977±0.003
SVM	0.9907±0.005	0.9961±0.004	0.9981±0.003	0.9985±0.002
RF	0.9876±0.006	0.9970±0.003	0.9968±0.004	0.9968±0.003

For RF, it can be seen that proposed FS model produce the highest accuracy for 10 subset of genes, i.e., 94%, and also perform better than existing approach for other subset of genes. In terms of kNN and SVM, the accuracy results of proposed FS model gives better results particularly for kNN, while for SVM it perform not well than HMM FS method, i.e., GFS-HMM produce high accuracy on 5 and 20 subsets of genes but produce worse accuracy on 10, 15, 25 and 30 genes. In addition, we describe the prediction accuracy of both FS models using ROC curves and Area Under the Curve (AUC) values. It is well known that a model has better predictive ability if the area between its ROC curve and the diagonal is larger, or we can say that the larger the area, the more separation the classes and has better prediction ability. Also, the higher the AUC value indicates its better classification or prediction ability.

Furthermore, the performance of individual FS models is depicted in Table A.1 in Appendix A. Clearly, it can be seen that all three models perform poorly as compared to stochastic FS models.

Fig. 3 displays ROC curves and AUC values of considered FS methods for three classifiers. In Fig. 3(a) we see, a ROC curve and AUC values for kNN, in which proposed model gives a good separation between both classes, i.e., normal and tumor, and has higher AUC of 0.86 as compared to 0.78 of HMM model. Similarly, for SVM and RF, Figs. 3 (a) and (c) shows the ROC curves and AUC values for considered FS models. It can be seen that GFS-HMM outperform the HMM model in terms of prediction ability using ROC curve and AUC. In addition, precision plots for HMM and GMFS-HMM under three classifiers have been displayed in Figure B1 in the Appendix B.

4.4 Biological interpretation of selected genes

In this section, we analyzed the biological relevance and previous studies history of the selected genes by proposed methodology. After applying proposed approach, 15 genes are selected from the colon dataset as shown in Table 2. In this Table, we have presented the rank position, ID and description of the chosen genes. Genes that are selected after proposed approach are highly significant and related to cancerous tissues. From previous studies [47, 48], the gene T71025 is identified to be the differentially expressed gene in colon cancer data. It is in chromosome 9q34.11 with pathway proteasome/UPS/NF-κ B/ cell-cycle. Moreover, T54364 is a dysregulated cell and expressed in tumor proliferation, mutation, and chemotherapy resistance; thus, it is an important pathway-level factor of colon cancer progression. Hence, this gene is responsible for degradation of

cellular protein. Furthermore, the selected genes with GenBank accession numbers H08393, H43887, M63391 and X57346 have been found and verified to be biologically downregulated in colon cancer patients. These were also chosen in the previous research [47, 49–52]. Similarly, we can also observe that gene U30872 is significant with previous work [53], with major pathways, JAK-STAT, ERK, and NF-κB. This gene is regulated for immune system regulation and stimulates tumor progression and metastasis. In summary, the HMM model based on multivariate feature selection techniques for genes selection identified genes in colon data that have been chosen in existing relevant works.

Example 2 **(RNA-seq PANCAN dataset)**: In this section, we analyze TCGA Pan-cancer Hiseq data [54] contains five types of cancer names: kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), Breast invasive carcinoma (BRCA), Prostate adenocarcinoma (PRAD) and lung adenocarcinoma (LUAD). It has 801 instances and 20531 features; where BRCA, COAD, KIRC, LUAD and PRAD contain 300, 78, 146, 141 and 136 instances, respectively. Our main aim is to assess proposed model that find the best performance while classifying 5 cancer types using accuracy as an evaluation metric.The detailed biological and functional interpretation of each gene is beyond the scope of this study. Furthermore, the data is preprocessed and normalized to streamline the analysis and improve the experimental effects.

Initially, 50% features (10131 out of 20263) were chosen by three filters including, IG, GI and mRMR. Then, based on prior works, we set the number of top most selected features by proposed methodology as {50, 100, 150, 200} from the top most 10131 features set. Table 3 presents the mean±SD of the accuracy results computed by three classifiers for GMFS-HMM. It shows that the proposed approach (GMFS-HMM) yields best performance with all classifiers (overlapping among intervals) and classification accuracy lies between 97% to 100% in all aspects. The accuracy results are slightly increased (up to 2%) whenever the selected features are larger than 50.

Recently, one of the main challenges, in gene expression data mining, is the problem of imbalanced data [55] classification which is prevalent in real-world applications. It is due to wide differences in the number of class instances or samples, and classifiers tend to favor the majority class and ignore the minority class prediction. The TCGA Pan-cancer data is comprised of imbalanced number of samples, particularly; class COAD is exactly 26% samples of class BRCA with high imbalanced ratio 3.85. We therefore employ Synthetic Minority Oversampling Technique (SMOTE) [55] to deal with imbalanced data problems. The average test-set accuracy and standard deviation are reported in Table 4.

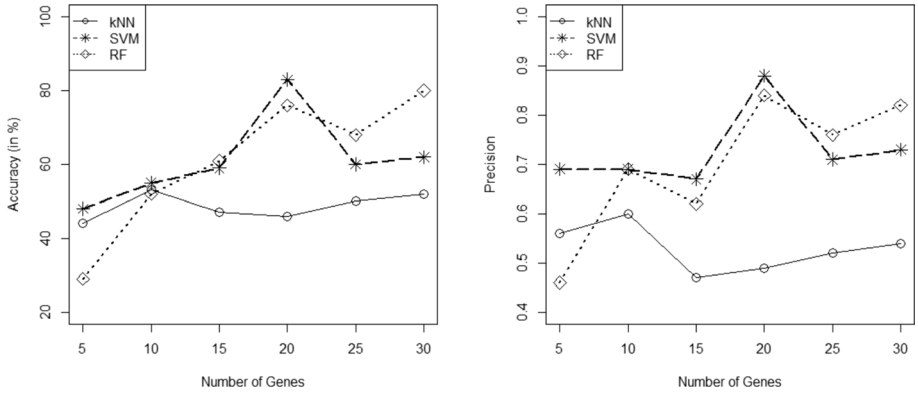

Fig. 4 Accuracy and precision results at different selected genes by proposed method on SRBCT data

Table 5 Accuracy results of proposed methodology at different selected subset of proteins for two groups of Mice data

	Control Mice			Trisomic Mice		
Proteins	***kNN***	***SVM***	***RF***	***kNN***	***SVM***	***RF***
10	0.327	0.368	0.443	0.438	0.546	0.636
20	0.195	0.468	**0.583**	0.478	0.539	0.545
30	0.511	0.46	0.500	0.408	0.559	0.545
40	0.437	0.466	0.417	0.429	0.58	0.364
50	0.362	0.475	0.333	0.450	0.565	0.636
60	0.543	0.479	0.416	0.45	0.528	0.636
70	0.458	0.478	0.500	0.451	0.548	**0.727**

Table 6 Learning information of mice data

No	Cluster Name	Group	Number of Mice
1	Normal	Control	19
2	Failed	Trisomic	7
3	Rescued	Trisomic	9
4	No learning	Control + Trisomic	37

Note that we adjust the sample size of each type of cancer to 300, and it resulted in increase of total sample size to 1500.

We see clearly that GMFS-HMM provide best accuracy for all subset of chosen features, but SVM perform slightly better than kNN and RF. From Table 4, we can also observe that if subset is higher than 50 there is some increasing trend in accuracy but not significantly different; overall, the accuracy is slightly improved due to SMOTE (as in non-SMOTE minimum and maximum accuracies are 0.9775 and 0.9960 respectively, while in SMOTE these results are 0.9832 and 0.9985, respectively). Furthermore, F_1 -score along with accuracy was also computed and presented in the Appendix A and Table A.2 for 100 selected features, and results indicate the overestimation of all three classifiers for 100 selected features. In summary, proposed hybrid-filter model together with kNN, SVM and RF classifiers can give best classification performance to the RNA-seq Pan-cancer data.

Example 3 **(SRBCT Data)**: For further illustration of GMFS-HMM, we investigated a multi-class gene expression data, named Small Round Blue Cell Tumors (SRBCTs). This data was originally studied by [56], and also available in R package **plsgenomics**. The SRBCT data contains expression levels of 2308 genes collected from glass-slide cDNA microarray, 83 samples (or patients) and 4 types (or classes) of cancer.

As we discussed in section 1, the major drawback of HMM based feature selection model is infeasibility for multi-class data. Therefore, its results are not presented in the Fig. 4. As the results in left penal of Fig. 4 shows, proposed model has obtained the highest accuracy, i.e., 83% for SVM and 20 chosen subset of genes among the all aspects. In contrast, the RF achieves reasonable accuracy performance for 20 to 30 selected genes, while kNN performs the worst among the three classifiers. Moreover, under the three classifiers, in right penal of Fig. 4, we display the prediction results using precision metric instead of ROC. The reason is that in multi-class problems, the ROC results are not

useful [57], and the authors suggest that precision is a more effective metric in such problems and ability to minimize False Positives across classes. From Fig. 4, the similar trends in precision results can be observed as we discussed for accuracy. The maximum precision is obtained for SVM with 20 selected genes.

Nevertheless, SRBCT molecular profiles are biologically important because it gives real and meaningful information about disease (cancer or tumor). After applying the proposed approach, a subset of 20 selected genes classified the SRBCT tumor types efficiently. These 20 chosen gene signatures provide biological distinct profiles that enhance the genetic interpretation and reduce classification error in clinical diagnostics (for detail, see [56]).

Example 4 **(Protein mice data)**: Down syndrome (DS) or Alzheimer named after English Dr. John Langdon Down, who diagnose this condition in people. Dr. Jerome Lejeune find that DS is a genetic abnormality in which a person has three copies of chromosome 21, instead of two, which is unnecessary genetic chromosomes causes problem in the normal growth of brain and body parts. DS badly affects the cognitive function which is associated with remembering, attention and thinking, decision making and learning. DS is due to genetic cause as it occurs due to extra copy of chromosome. It is frequent in humans as 1 in 700 births in USA and 1 in 1000 worldwide [58].

In this section, the proposed model is applied to the expression of 77 protein obtained from the nuclear -enriched fraction of cortex from control and trisomic mice [59]. This dataset contains 77 expression levels of proteins. There were two types of mice that are, 38 control mice and 34 trisomic mice as 72 of total mice as samples subjects. The dataset is consisting of 8 classes. These classes are c-CS-s, c-CS-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m, c-SC-s, and c-SC-m. In each class, the number of mice is ranges from 7 to 10. There are replicates of each protein level to obtain the protein for each class, i.e., proteins for class c-CS-m. We take average of 15 replicates for each protein for this purpose to obtain the original data matrix with dimension 72×77, where, n=72, and p=77.

To discriminate the important proteins, the analysis is performed on control mice and trisomic mice separately. First, the proposed filter model is applied on the control mice data for the selection of subset of proteins. This part of mice data contains 77 proteins of 38 mice. Secondly, the trisomic mice are analyzed with n= 34 subjects and same number

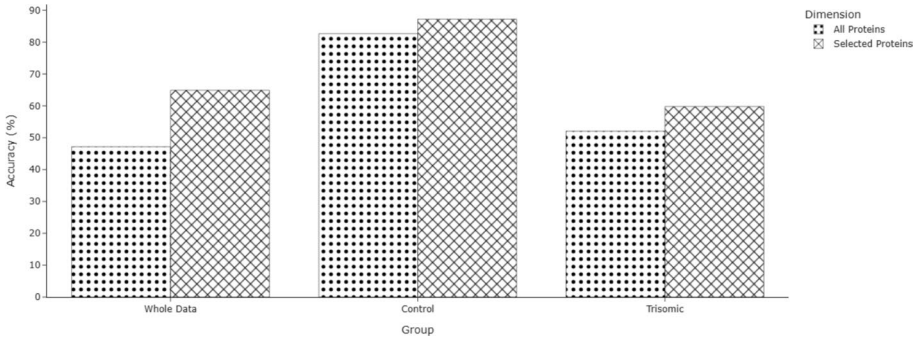

Fig. 5 Distribution of original and reduced datasets prediction accuracy across different mice groups. The data is reduced by proposed hybrid model

of proteins. This analysis will enhance our understanding of reliable classification of proteins between control and trisomic groups. The results are depicted in Table 5.

For control group, it can be observed that RF provides the best prediction accuracy for 20 chosen proteins in comparisons with other subsets of proteins, i.e., 58%. This finding suggests that the data do not contain more proteins, near to 77, associated with control group. On the other hand, for trisomic group, again the highest accuracy (i.e., 73%) is produced by RF for chosen subset size 70, indicating that most of the protein in the data are liable for trisomic disease in mice. Note that the accuracy results are provided only for the proposed model due to the fact that the existing HMM selection method was not applicable because of the multi-class nature of mice data.

4.5 Protein set enrichment analysis

In this part, we conduct an experiment to detect four Alzheimer's disease pathways or learnings in the mice data including normal, rescued, failed and no learning with deferentially expressed proteins. According to the authors [59], proteins in mice data plays an important role in the brain development structure and function. Our aim is to evaluate the proposed model for clustering mice into four leaning based on selected expression of proteins. A summary of the mice data content obtained from [59] is presented in Table 6. However, the complete mice data contains 77 proteins include 72 sample patients from the control and trisomic groups. From biological point of view, it is important to note that processing the data for individual groups is more crucial than complete mice data [59]. Therefore, we conduct an analysis on the entire dataset as well as for control and trisomic parts. Furthermore, due to the better performance of SVM on this data, we consider only SVM for computing classification accuracy.

Figure 5 displays classification accuracy and optimum subset of proteins for the proposed model of each part of data. As can be seen from Fig. 5, the protein subsets chosen for all three datasets by new hybrid model achieve higher accuracies compared to datasets without protein selection. For the chosen optimal subsets of 30, 10 and 40 proteins, the increases in clustering accuracy are computed as 17%, 5% and 7% for the whole, control and trisomic datasets, respectively. This suggests the effectiveness of the new model in efficiently classifying Alzheimer's patients to their true learning pathways in the mice protein data.

In the future, we plan to explore multivariate FS methods including PLS-based methods [61, 62] and [63] for dealing with redundancy issues and improving the performance of the proposed model.

5 Sensitivity analysis of parameter ϕ

For a fixed subset size of 20 selected features, Table 7 presents the evaluation metrics for different values of the ϕ in the proposed GMFS-HMM model. The results indicate that the proposed model achieves the maximum accuracy when ϕ is approximately 2. Additionally, the precision peaks are obtained at $\phi = 1.5$ and $\phi = 3$. Based on these observations, it is recommended that the value of ϕ is close to 2 to obtain optimal prediction performance with the proposed GMFS-HMM model.

Table 7 Prediction performance across varying values of the ϕ parameter for the proposed GMFS-HMM model using Colon dataset

ϕ	Classifier	Sensitivity	Accuracy	Specificity	Precision
1	kNN	0.862	0.678	0.386	0.709
	SVM	0.868	0.679	0.27	0.723
	RF	0.926	0.675	0.243	0.691
1.5	kNN	0.885	**0.755**	0.534	**0.779**
	SVM	0.871	0.666	0.305	0.668
	RF	0.934	**0.738**	0.423	0.739
2	kNN	0.871	0.667	0.351	0.69
	SVM	0.851	0.572	0.189	0.588
	RF	0.927	0.673	0.225	0.690
3	kNN	0.864	0.682	0.391	0.711
	SVM	0.871	0.724	0.312	**0.7862**
	RF	0.929	0.651	0.206	0.67
3.5	kNN	0.874	0.687	0.376	0.715
	SVM	0.821	0.48	0.172	0.467
	RF	0.924	0.667	0.219	0.68
4	kNN	0.871	0.678	0.367	0.708
	SVM	0.877	0.681	0.257	0.721
	RF	0.914	0.655	0.235	0.684

6 Conclusions

This study designed a hybrid FS model by unifying three various multivariate FS filter models including: mRMR, IG and GI in the framework of HMM. We adopt this methodology to work with both binary and multi-class high-dimensional genes data. Due to the use of multivariate filter models, it becomes able to take into account the problem of redundancy among genes with low computational cost. The performance of selected genes subsets are assessed through kNN, SVM and RF classifiers. For binary colon data, our findings demonstrated that the proposed FS approach performs better than the HMM FS method in terms of informative genes (see Table 1 and Fig. 3). The GMFS-HMM has also been evaluated on multi-class SRBCT data and yield the best accuracy for SVM with 20 selected subset of genes (see Fig. 4). Furthermore, the results on RNA-seq TCGA Pan-cancer dataset show the superiority of new approach GMFS-HMM evaluated by three well-known machine learning classifiers. The analysis of Alzheimer's data further supports the application of proposed methodology for identifying informative proteins and predictions. Moreover, GMFS-HMM selected significant genes or proteins that are consistent with the state-of-the-art methods on tumor classification and efficient clustering of four learning pathways in mice data.

Additional Results

See Table 8.

The Table 9 shows the average F1-score along with average accuracy and ± Standard Deviation (SD) results of proposed methodology GMFS-HMM for 100 selected features. The results show that all three classifiers perform a little bit over optimistic in terms of accuracy and F1-score in case of an imbalanced dataset. The best findings were obtained for SVM in aspects of imbalanced datasets while kNN performed well in imbalanced

Table 8 Accuracy of separate FS models: mRMR, IG and GI for colon data

	mRMR			IG			GI		
Genes	**kNN**	**SVM**	**RF**	**kNN**	**SVM**	**RF**	**kNN**	**SVM**	**RF**
5	0.347	0.42	0.415	0.475	0.359	0.343	0.412	0.368	0.435
10	0.52	0.486	0.4	0.527	0.483	0.532	0.356	0.384	0.403
15	0.332	0.448	0.381	0.548	0.481	0.511	0.324	0.328	0.352
20	0.372	0.456	0.28	0.218	0.34	0.301	0.349	0.32	0.281
25	0.344	0.396	0.36	0.392	0.329	0.37	0.288	0.48	0.508

Table 9 Results comparison of three classifiers on RNA-seq data in terms of average F_1-score and Accuracy ± SD

	Without SMOTE		SMOTE	
Classifier	F_1***-Score***	**Acc ± SD**	F_1***-Score***	**Acc ± SD**
kNN	0.994	0.994 ± 0.004	0.975	0.971 ± 0.008
SVM	0.990	0.993 ± 0.005	0.988	0.982 ± 0.007
RF	0.986	0.984 ± 0.007	0.978	0.983 ± 0.010

data. Therefore, balancing through SMOTE technique is important before implementing the proposed model to imbalanced data.

Precision plots for Colon data

See Fig. 6.

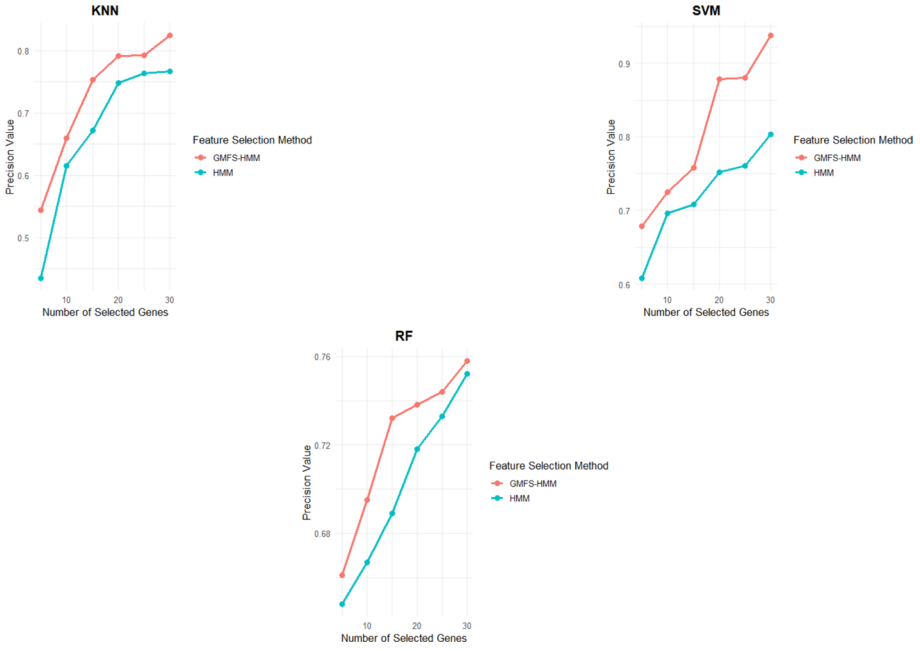

Fig. 6 Precision plots for different gene subsets on Colon data

Author contributions
A.W. and S.H. conceptualized the research, methodology, analysis and software. A.W.: validation and supervision. S.H.: Writing initial draft. U.A.: Data curation and analysis. W.H. and M.A.: visualization, review and editing.

Funding
No external funding supported this project.

Data availability
"Data availability statement The datasets that support the findings of this study are available from the following sources: Colon and RCBCT Datasets: Openly available in (https://cran.r-project.org/web/packages/plsgenomics/plsgenomics.pdf; http://genomicspubs.princeton.edu/oncology/). Mice Data: All relevant data are available via Figshare (http://dx.doi.org/10.6084/m9. figshare.1421985). RNA-seq PANCAN dataset: This dataset is publicly available on (https://www.kaggle.com/datasets/waalbannyantudre/gene-expression-cancer-rna-seq-donated-on-682016)."

Declarations

Consent to publish
Not applicable

Ethical approval
Not applicable

Consent to participate
Not applicable

Conflict of interest
The authors declare no conflict of interest.

Received: 6 August 2025 / Accepted: 20 April 2026
Published online: 05 June 2026

References

1. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine, Database, vol. 2020, pp. baaa010, 2020.
2. Donoho DL. High-dimensional data analysis: The curses and blessings of dimensionality. AMS math challenges lecture. 2000;1(2000):32.
3. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293–314.
4. Jia W, Sun M, Lian J, Hou S. Feature dimensionality reduction: a review. Complex Intel Syst. 2022;8(3):2663–93.
5. Arowolo MO, Isiaka RM, Abdulsalam SO, Saheed Y, Gbolagade KA. A comparative analysis of feature extraction methods for classifying colon cancer microarray data, EAI endorsed transactions on scalable information systems, 2017;4:14, .
6. Liu Y. Feature extraction for DNA microarray data. In: Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07) 2007;371-376.
7. Alhenawi Ea, Al-Sayyed R, Hudaib A, Mirjalili S. Feature selection methods on gene expression microarray data for cancer classification: A systematic review, Computers in biology and medicine, 2022;140:105051.
8. Khan DM, Yaqoob A, Iqbal N, Wahid A, Khalil U, Khan M, Abd Rahman MA, Mustafa MS, Khan Z. Variable Selection via SCAD-Penalized Quantile Regression for High-Dimensional Count Data, IEEE Access, 2019;7:153205-153216.
9. Wahid A, Khan DM, Iqbal N, Khan SA, Ali A, Khan M, et al. Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou's 5-steps rule. Chemom Intell Lab Syst. 2020;199:103958.
10. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
11. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform. 2015;2015(1):198363.
12. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE. 2015;1200–5
13. Bommert A, Welchowski T, Schmid M, Rahnenführer J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data, Briefings in Bioinformatics, 2022;23(1): bbab354.
14. Del Moral P, Nowaczyk S, Pashami S. Why is multiclass classification hard? IEEE Access. 2022;10:80448–62.
15. Momenzadeh M, Sehhati M, Rabbani H. A novel feature selection method for microarray data classification based on hidden Markov model. J Biomed Inform. 2019;95:103213.
16. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinf. 2012;9(4):1106–19.
17. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 2020;143:106839.
18. Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). 2003;856–63.
19. Fox RJ, Dimmic MW. A two-sample Bayesian t-test for microarray data. BMC Bioinformatics. 2006;7:1–11.

20. Varma S, Simon R. Iterative class discovery and feature selection using Minimal Spanning Trees. BMC Bioinformatics. 2004;5:1–9.
21. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005;21(5):631–43.
22. Bø TH, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biol. 2002;3:1–11.
23. Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics. 2006;22(14):e184–90.
24. Meyer PE, Schretter C, Bontempi G. Information-theoretic feature selection in microarray data using variable complementarity. IEEE J Select Top Signal Process. 2008;2(3):261–74.
25. Lee I-H, Lushington GH, Visvanathan M. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. J Clin Bioinf. 2011;1:1–8.
26. Cao Z, Wang Y, Sun Y, Du W, Liang Y. A novel filter feature selection method for paired microarray expression data analysis. Int J Data Min Bioinform. 2015;12(4):363–86.
27. Munirathinam DR, Ranganadhan M. A new improved filter-based feature selection model for high-dimensional data. J Supercomput. 2020;76(8):5745–62.
28. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. J Biomed Inform. 2018;85:189–203.
29. Zhou H, Zhang J, Zhou Y, Guo X, Ma Y. A feature selection algorithm of decision tree based on feature weight. Expert Syst Appl. 2021;164:113842.
30. Obaido G, Achilonu O, Ogbuokiri B, Amadi CS, Habeebullahi L, Ohalloran T, Chukwu CW, Mienye E, Aliyu M, Fasawe O. An improved framework for detecting thyroid disease using filter-based feature selection and stacking ensemble, IEEE Access, 2024.
31. Gong H, Li Y, Zhang J, Zhang B, Wang X. A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information. Eng Appl Artif Intell. 2024;131:107865.
32. Mishra V, Sharma V, Mishra U. A hybrid approach for leaf classification using machine learning and deep learning. In: 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT). IEEE. 2023;1589–93. https://doi.org/10.1109/ICCPCT58313.2023.10245548.
33. Wang Y, Feng L. A new hybrid feature selection based on multi-filter weights and multi-feature weights. Appl Intell. 2019;49:4033–57.
34. Mining WID. Data mining: Concepts and techniques. Morgan Kaufinann. 2006;10(559–569):4.
35. Verma S, Sahu SP, Sahu TP. Two-stage hybrid feature selection approach using levy's flight based chicken swarm optimization for stock market forecasting. Comput Econ. 2024;63(6):2193–224.
36. Kumar JM, Rauf HA, Umamaheswari R. Switched capacitor-coupled inductor DC–DC converter for grid-connected PV system using LFCSO-based adaptive neuro-fuzzy inference system. J Circuits Syst Comput. 2020;29(12):2050201.
37. Wei G, Zhao J, Feng Y, He A, Yu J. A novel hybrid feature selection method based on dynamic feature importance. Appl Soft Comput. 2020;93:106337.
38. Mohtashami M, Eftekhari M. A hybrid filter-based feature selection method via hesitant fuzzy and rough sets concepts. Iran J Fuzzy Syst. 2019;16(2):165–82.
39. Xing EP, Jordan MI. R. Feature selection for high-dimensional genomic microarray data: M. Karp; ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning2001. p. 601–8.
40. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005;3(02):185–205.
41. Shang W, Huang H, Zhu H, Lin Y, Qu Y, Wang Z. A novel feature selection algorithm for text categorization. Expert Syst Appl. 2007;33(1):1–5.
42. Liu H, Zhou M, Lu XS, Yao C. Weighted Gini index feature selection method for imbalanced data. In: 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE; 2018;1-6. https://doi.org/10.1109/ICNSC.2018.8361371.
43. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat. 1970;41(1):164–71.
44. Collier N, Nobata C, Tsujii Ji. Extracting the names of genes and gene products with a hidden Markov model 2000.
45. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
46. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci. 1999;96(12):6745–50.
47. Apiletti D, Baralis E, Bruno G, Fiori A. Maskedpainter: feature selection for microarray data analysis. Intelligent Data Analysis. 2012;16(4):717–37.
48. Chen JJ, Tsai C-A, Tzeng S, Chen C-H. Gene selection with multiple ordering criteria. BMC Bioinformatics. 2007;8:1–17.
49. Alladi SM, Ravi V, Murthy US. Colon cancer prediction with genetic profiles using intelligent techniques. Bioinformation. 2008;3(3):130.
50. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. Genome Res. 2001;11(11):1878–87.
51. Wessels LF, Reinders MJ, van Welsem T, Nederlof PM. Representation and classification for high-throughput data. In: biomedical nanotechnology architectures and applications. SPIE. 2002;4626:226–37
52. Yap Y, Zhang X, Ling M, Wang X, Wong Y, Danchin A. Classification between normal and tumor tissues based on the pairwise gene expression ratio. BMC Cancer. 2004;4:1–17.
53. Kim K-J, Cho S-B. Prediction of colon cancer using an evolutionary neural network. Neurocomputing. 2004;61:361–79.
54. Nitta Y, Borders M, Ludwig SA. Analysis of gene expression cancer data set: Classification of TCGA pan-cancer HiSeq data. In: 2021 IEEE International Conference on Big Data (Big Data). IEEE. 2021;4745-4752.
55. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–57.
56. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673–9.
57. Aguilar-Ruiz JS, Michalak M. Multiclass classification performance curve". IEEE Access. 2022;10:68915–21.

58. Hospital BC. Down syndrome. Retrieved from. https://www.childrenshospital.org/conditions/downsyndrome 300 Longwood Avenue Boston, MA 02115; 2024.
59. Higuera C, Gardiner KJ, Cios KJ. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. PLoS ONE. 2015;10(6):e0129126.
60. Liu H, Zhou M, Lu X, Yao C. Weighted Gini index feature selection method for imbalanced data, 2018 IEEE 15th ICNSC, 1-6.
61. You W, Yang Z, Ji G. PLS-based gene subset augmentation and tumor-specific gene identification. Comput Biol Med. 2024;174(4):10843.
62. You W, Yang Z, Ji G. PLS-based recursive feature elimination for high-dimensional small sample. Knowl-Based Syst. 2014;55:15–28.
63. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1):389–422.
64. Yang QT, Xu XX, Zhan ZH, Zhong J, Kwong J, Zhang J. Evolutionary multitask optimization for multiform feature selection in classification. IEEE Trans Cybernet. 2025;55(4):1673–86.
65. Ding J, Du J, Wang H, Xiao S. A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning. Sci Rep. 2025;15(1):16828.
66. Wang X, He Q, Jian W, Meng H, Zhang B, Jin H, et al. Hybrid feature ranking and classifier aggregation based on multi-criteria decision-making. Expert Syst Appl. 2024;238:122193.
67. Chaabene S, Boudaya A, Bouaziz B, Chaari L. An overview of methods and techniques in multimodal data fusion with application to healthcare. Int J Data Sci Anal. 2025;20(4):3093–117.
68. RoselinKiruba R. A hybrid machine learning framework for early hepatitis detection using enhanced feature selection and ensemble classification. nt. j. inf. tecnol. 2026; 1-11
69. Sharma N, Dutta M. Dynamic Ensemble Learning in Recommendation Systems: A Comprehensive Review. Journal of Information & Knowledge Management. 2026: 2550133

Publisher's Note