A Novel Hybrid Approach for Chronic Disease Classification

Divya Jain, Computer Science and Engineering, The NorthCap University, Gurugram, India
Vijendra Singh, Computer Science and Engineering, The NorthCap University, Gurugram, India

ABSTRACT

A two-phase diagnostic framework based on hybrid classification for the diagnosis of chronic disease is proposed. In the first phase, feature selection via ReliefF method and feature extraction via PCA method are incorporated. In the second phase, efficient optimization of SVM parameters via grid search method is performed. The proposed hybrid classification approach is then tested with seven popular chronic disease datasets using a cross-validation method. Experiments are then conducted to evaluate the presented classification method vis-à-vis four other existing classifiers that are applied on the same chronic disease datasets. Results show that the presented approach reduces approximately 40% of the extraneous and surplus features with substantial reduction in the execution time for mining all datasets, achieving the highest classification accuracy of 98.5%. It is concluded that with the presented approach, excellent classification accuracy is achieved for each chronic disease dataset while irrelevant and redundant features may be eliminated, thereby substantially reducing the diagnostic complexity and resulting computational time.

KEYWORDS

Chronic Disease Diagnosis, Hybrid Approach, Optimized Support Vector Machines, PCA, Relief

1. INTRODUCTION

Lately, the devastating effects of chronic diseases on the health and wellbeing of populations all over the globe are becoming worrisome. In 2016, the World Health Organization (WHO) estimates a total of 383,000 deaths in 2015 arising from asthma with 235 million people suffering from this debilitating chronic illness that commonly affects children (WHO, 2016). Also, in 2015, the WHO (2016) statistics reveal that around 13.4 lakh deaths occurred due to viral hepatitis. More recently, in 2017, WHO reports that approximately 205 million women suffer from diabetes worldwide (WHO, 2017). Moreover, these effects are likely to be exacerbated with the rapidly aging populations growing in many countries all over the world. Conceivably, it is essential, in times such as these, that efficient diagnostic mechanisms be developed to render best possible remedy to patients inflicted with various chronic diseases in the least amount of time to sustain a decent level of population health and wellbeing in the developed, developing and under-developing countries of the world.

Accordingly, early and accurate diagnosis of chronic diseases plays a key role in medical informatics applications and clinical care services. With the advent of the digital era, data analytics in the medical field have continued to grow in handling voluminous amount of patient generated data with multidimensional characteristics. Big data mining provides efficient tools and techniques that can aid in health insurance decision-making (Tan et al., 2009) and other clinical decision-making (Han et al., 2011; Larose & Larose, 2014). As medical datasets contain all sorts of data from heterogeneous
sources, there is an utmost need to develop innovative analytical methods and to formulate effective learning algorithms to deal with high-dimensional data.

The rest of this paper is organized as follows. Section 2 discusses the background on existing classifiers for high data dimensionality reduction and argues the fundamental basis for undertaking our proposed approach for data analytical innovation. Section 3 details the dataset materials and methods used in this research. Section 4 depicts the hybrid model being proposed. Following this, Section 5 presents the study results and their comparative evaluation vis-à-vis other analytical methods. Section 6 summarizes the findings, highlighting key effects of the various statistical significance tests. Finally, Section 7 concludes the paper by reviewing the study contributions, provides insights on implications of the findings and potential future work for others interested in this line of research.

2. BACKGROUND ON EXISTING CLASSIFIERS

Over the years, numerous machine-learning methods have been applied successfully for effective disease diagnosis (Azzawi et al., 2016; Dora et al., 2017). Owing to its many excellent features and remarkable generalization performance (Cortes & Vapnik, 1995; Critianini & Shawe-Taylor, 2005), support vector machine (SVM) classifier is a widely used learning method in huge demand today. SVM is based on finding the maximum-margin hyper-plane for the separation of two classes as wide as possible. It has gained popularity in a wide range of biological applications (Noble, 2006). Also, with the use of kernel functions, SVM works efficiently with linear as well as non-linear datasets.

With the optimization of parameters used in kernel functions, the SVM classification accuracy increases at a significant rate. Grid search technique, a widely used method for optimizing SVM classifier, helps in finding best parameters to tune the performance of SVM (Lin et al., 2008; Tang et al., 2009). Although SVM has exceptional advantages, its performance will significantly improve if dimensionality reduction techniques are applied before the classification of data (Xie et al., 2010). Notably, researchers use feature selection extensively to reduce high data dimensionality (Dash & Liu, 1997; Guyon & Elisseeff, 2003; Chizi & Maimon, 2009; Kumar & Minz, 2014; Ang et al., 2015; Shardlow, 2016). For disease diagnosis, feature selection eliminates the attributes that are least significant to a particular disease.

The extant literature cumulates studies demonstrating noteworthy results from authors who have conducted research in the field of high data dimensionality reduction with classification. Polat & Gunes (2007), for example, proposed a novel medical system based on principal component analysis (PCA), fuzzy approach and adaptive neuro-fuzzy inference system (ANFIS) classifier for the diagnosis of lymph diseases, giving promising results with a high classification accuracy of 88.83%. Babaoğlu et al. (2010) investigated the effect of PCA-SVM based approach for the diagnosis of coronary artery disease (CAD). Their experimental analysis and results demonstrated that the PCA-SVM system not only diminishes the high data dimensionality and training-test timing, but also increases the accuracy of the SVM classifier.

More recently, Pang et al. (2015) applied ReliefF-SVM based method for the computer-aided diagnosis of breast tumor, yielding positively appealing results with 90.0% accuracy rate, 98.7% sensitivity and 73.8% specificity rate. Shah et al. (2017) employed SVM classifier with probabilistic PCA through parallel analysis for the automatic detection of heart disease. Statistically, their results showed that the presented method reduced approximately 70% of the dimensionality of the dataset with highest classification accuracy of 91.30%. Sangaiah & Kumar (2018) presented a hybrid approach using ReliefF method and entropy-based genetic algorithm for the diagnosis of breast cancer. The results they obtained depicted a superior performance of the approach with reduced dimensionality and significant accuracy rate. Finally, Singh & Pandey (2018) proposed an efficient diagnostic system using a unified two-phase approach based on PCA and K-nearest neighbor classifier (PCA-KNN) for the detection of liver disease. Their prediction model attained remarkable results in terms of various evaluation measures such as positive v. negative predictive value, accuracy, specificity and sensitivity.
Altogether, the referenced statistics above show that, for medical diagnosis, both the PCA and ReliefF approaches have exhibited good performance individually with a SVM classifier. Motivated by these results, the present study aims to evaluate the application of a hybrid approach where both the PCA and ReliefF are combined with SVM classifier for achieving data dimensionality reduction and data classification improvements in disease prediction. Specifically, a machine-learning framework comprising two operational phases will be employed: Phase-I emphasizes the identification of the most crucial features from the disease dataset via a combination of ReliefF-PCA method applied with specific threshold. The resultant first-phase dataset with the reduced set of features will then be passed onto Phase II of the analysis as Phase-II does efficient optimization of SVM parameters and finds the best values of SVM parameters, namely, ‘cost’ and ‘gamma’.

Details of the dataset materials and methods applied in this research will now be discussed.

3. MATERIALS & METHODS

3.1. Dataset Description

This research encompasses seven (7) sets of disease data for which the proposed machine-learning framework is being applied, following which the approach is further evaluated on the basis of various performance measures. Both large and small sized datasets are also being considered for the experimentation. Table 1 provides the details of these various chronic disease datasets.

A detailed description of the separate chronic disease dataset is given below. It contains details about the number of samples, attributes and classes corresponding to each dataset.

- **Lung Cancer (Michigan) Dataset**: This dataset, obtained from biomedical repository (Ridge, 2005), contains 7129 features and 96 samples. Among them, 86 samples are classified as lung adenocarcinomas samples and 10 samples are classified as non-neoplastic lung samples.
- **Leukemia Cancer Dataset**: This dataset, acquired from online repository (Ridge, 2005), contains 7129 features. The total numbers of samples in the dataset are 72 out of which 47 are classified as ‘ALL’ and 25 samples are classified as ‘AML’.
- **Colon Cancer Dataset**: This is a high-dimensional dataset collected from biomedical repository (Ridge, 2005) comprising 2000 features and 62 samples. Among these, 40 tumor samples are classified as ‘negative’ while 22 normal samples are classified as positive.
- **Lung Cancer (Ontario) Dataset**: This dataset contains 2880 attributes and 39 instances. Among them, 24 tumor samples are classified as ‘relapse’ whereas 15 patients’ normal samples are classified as ‘non-relapse’.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total Features</th>
<th>Total Samples</th>
<th>Type</th>
<th>Number of classes</th>
<th>Source/ Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer (Michigan)</td>
<td>7129</td>
<td>96</td>
<td>Microarray</td>
<td>2</td>
<td>Kent-Ridge Biomedical Dataset Repository</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>2000</td>
<td>62</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Leukemia Cancer</td>
<td>7129</td>
<td>72</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Lung Cancer (Ontario)</td>
<td>2880</td>
<td>39</td>
<td>Text</td>
<td>2</td>
<td>UCI Machine Learning Repository</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>25</td>
<td>400</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Hepatitis</td>
<td>19</td>
<td>155</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Heart Disease</td>
<td>13</td>
<td>270</td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
3.2. Pre-processing of Data

To improve the quality of big data mining and to remove missing values and redundant data, data pre-processing is applied on all seven datasets a priori, that is, before applying the proposed machine-learning framework.

Researchers have realized that good data preparation and effective data pre-processing is essential in order to use machine-learning algorithms and big data mining tools on datasets effectively (Pyle, 1999; Ilango & Ramaraj, 2010). Moreover, in medical informatics, the diagnosis of disease also becomes quicker and easier if data is free from missing, redundant and irrelevant data. The preliminary analysis of some datasets indicates the presence of some missing values; hence, for numerical data, mean across each column containing the missing values is computed and that value replaces the missing data. For non-numerical data, the most frequently occurring value is identified as the mode across the column and that value replaces the missing data.

3.3. Integration of Feature Selection Techniques – ReliefF & PCA

Feature selection plays an integral role in many of the application areas that are related to expert and intelligent systems, such as data mining and machine learning (Gao et al., 2018). Choosing among the appropriate feature selection methods prior to applying any learning technique is important as it not only ensures high performance but also reduces cost incurred in data computation (Che et al., 2017; Peng & Fan, 2017).

First, learning algorithm is greatly affected by the ‘curse of dimensionality’ (Powell, 2007). Second, the presence of large number of features in a dataset also increases the search space. Third, with limited training examples, working with a huge number of features becomes non-trivial, as it would lead to over-fitting. In medical informatics, the curse of dimensionality can be predominantly found in high-dimensional datasets. If the original dataset is passed to the classifier, it will take too long and consume too much memory for computational processing; hence, it is essential to remove those unnecessary, undesired and/or insignificant features especially from those high-dimensional datasets.

For disease prediction, it is essential to identify the most critical parameters related to a particular disease (Jain & Singh, 2018). Variable selection eliminates redundant and irrelevant parameters of the medical datasets; in turn, this would hasten the computational time for the classifier to provide accurate results. In healthcare big data analytics, dimensionality reduction methods are utilized for the diagnosis of numerous chronic diseases such as diabetes, heart disease, hepatitis, chronic kidney disease, arthritis, leukemia, lung cancer, breast cancer and more (Lu et al., 2014; Chowdhary & Acharya, 2016; Polat et al., 2017).

In the past, researchers have utilized different algorithms for dimensionality reduction such as Relief (Kira & Rendell, 1992); ReliefF (Kononenko, 1994); correlation-based feature selection or CFS (Hall, 2000); fast correlation-based filter selection or FCBS (Yu & Liu, 2003); and FAST (Nagpal & Gaur, 2015; Song et al., 2013) to discover the most critical parameters relevant to a particular disease. Today, researchers are focusing on the development of new hybrid dimensionality reduction methods that accelerate the removal of useless and least important features and show better accuracy of classification in comparison to existing methods (Uzer & Yilmaz, 2013; Puyalnithi & Vankadara, 2018). This research presents ReliefF-PCA based hybrid method for the dimensionality reduction of disease datasets.

- **Chronic Kidney Disease Dataset**: This dataset, obtained from UCI repository (Lichman, 2013), contains 400 instances with multivariate characteristics. The total numbers of dimensions in it are 25. It has two classes: ‘ckd’ and ‘notckd.’
- **Heart Disease Dataset**: This multivariate dataset contains 13 features and 270 instances (Lichman, 2013). It contains both categorical and real values with class values indicating the presence and absence of the disease.
- **Hepatitis Dataset**: This dataset, acquired from online repository (Lichman, 2013), contains 155 integers, real and categorical samples with 19 attributes in the dataset.
3.4. Classification via Optimized SVM Classifier

SVMs are widely used kernel-learning algorithms for solving high-dimensional prediction problems (Vapnik, 1995). Compared to other classifiers, the SVM classifier has been found to exhibit superior generalization ability and scales well with linear as well as non-linear data. Moreover, the SVM classifier attains relatively robust pattern recognition performance when used with well-defined approaches commonly applied in statistical learning and optimization theory.

The main goal of SVM classifier is to find a hyper-plane that separates positive samples from negative samples with the largest margin. The margin is stated as the distance of the hyper-plane to the closest of the positive and negative classes (Platt, 1988). The equation for the separating function is given by:

\[ p^w + a = 0 \]

where ‘p’ depicts ‘weight vector’, ‘u’ depicts ‘input pattern’ and ‘a’ depicts ‘bias’.

The objective is to maximize the margin, which is depicted as:

\[ \text{Margin} = \frac{2}{\|p\|} \]

For linearly separable data, it is easy to find the optimal hyper-plane separating two classes. However, for non-separable problems, SVM uses ‘Kernel Functions’ for non-linear mapping to high dimensional space. There exist many kernel functions, for example, linear kernel function (LKF), polynomial kernel function (PKF), sigmoid kernel function (SKF), exponential radial basis kernel function (ERBKF), and Gaussian radial basis kernel function (GRBKF). Out of these various kernel functions, researchers have identified the radial basis (RB) function as the best kernel function (Lin & Liu, 2007; Wang et al, 2017). Even so, the challenge with this kernel expression is to seek the best values of penalty parameters – “cost” and “gamma.”

As noted, the performance of SVM classifier is largely affected by ‘cost’ and ‘gamma’ parameters. While the ‘gamma’ parameter states the effect of training examples, the ‘cost’ parameter governs the effect of support vectors. The values of these two parameters significantly affect the classification accuracy of training and testing data. In this research, grid search technique is applied to identify their best values (Akay, 2009; Xie & Wang, 2011). This method does efficient parameter optimization of SVM parameters through a grid space of cost and gamma values. The values which give best accuracy are picked up to train and test data. Earlier studies have numerous examples that utilize SVM classifier for the diagnosis of chronic diseases (Chen et al., 2012; Zangooei et al., 2014; Peker, 2016; Singh & Pandey, 2016).

4. PROPOSED MACHINE-LEARNING FRAMEWORK

The proposed framework comprises two phases. The first stage (Phase-I) of the model selects relevant features for the dataset via a hybrid approach based on ReliefF method along with principal component analysis. The second stage (Phase-II) is based on finding best parameters of SVM classifier via the grid search method.
4.1. Phase-I: Efficient Feature Selection

In this phase, the aim is to present an efficient hybrid feature selection and apply it on all considered chronic disease datasets. The irrelevant features from disease datasets are eliminated with the proposed ReliefF-PCA based hybrid feature selection technique.

First, the chronic disease dataset is loaded into the system. The next step is to pre-process the data to remove missing values and redundant data values. The missing values are substituted by either mean or mode of the column based on the other data in the column. For missing numerical values, it is replaced with the mean value of the column; otherwise, for missing categorical values, it is replaced with the mode value of the column. Then ReliefF feature selection is applied on pre-processed data with the resultant yielding a weight matrix. The weight matrix contains the weights corresponding to each feature. Next, appropriate threshold is applied to select only relevant features from the dataset.

In this article, the threshold ‘theta’ is taken as mean (W). The proposed model is tested with different thresholds such as Mean, Median & Standard Deviation. It was found that the presented model works best with the threshold ‘theta’, where, ‘theta’ equals mean (W) (Jain & Singh, 2018). By choosing among those with a large W value (i.e. those that exceed a threshold ‘theta’), the final selection of attributes is performed. The resultant creates a list of features (L) with the elimination of irrelevant features from the dataset. This list (L) is passed to PCA method to remove redundant features from the dataset, which would assign a score to each feature. Finally, the dataset with the application of presented hybrid approach is passed onto second phase of the research. Figure 1 shows the proposed model.

4.2. Phase-II: Fast Parameter Optimization via SVM Classification

As the dimension of the datasets has been reduced, and the data set with new features has been rebuilt, machine-learning techniques can now be practiced and assessed on the basis of various performance measures.

In this research, a SVM with efficient parameter optimization is applied on seven chronic disease datasets. The resultant dataset after applying ReliefF-PCA based hybrid feature selection approach is passed onto Phase II of the research in which the dataset is partitioned into two portions – the training v. the testing set. After that, the optimization of SVM is done to determine best values of ‘cost’ and ‘gamma’. The best parameters of kernel function of SVM are determined via grid search algorithm. Grid search method searches all possible combinations and makes a grid space of ‘cost’ and ‘gamma’. For grid search method, to execute with less computational time, the concept of machine workers executing in parallel has been applied. By means of this method, all the possible combinations of parameter values are evaluated and the best combination yielding maximum accuracy is retained. With the optimized SVM parameters, classification is applied on each chronic disease dataset and assessed on the basis of accuracy.

Figure 2 shows the pseudo-code of the Phase-I & Phase-II followed for the proposed framework.

5. EXPERIMENTAL RESULTS

The proposed model is validated using the cross-validation method and evaluated on the basis of the features selected by the algorithm, the accuracy produced by the classifier and the time required to execute the algorithm. Of the seven (7) datasets studied, four (4) are large-sized while three (3) are small-sized datasets. The classification performance of the proposed system is also evaluated at the end of proposed feature selection via the ReliefF-PCA method. The presented classification method using optimized SVM is also compared with other existing classifiers that are applied also on the same chronic disease datasets.
5.1. Evaluating the Effect of Proposed Hybrid Feature Selection Approach on the Classifier Performance

This section evaluates the effect of proposed feature selection approach on the performance of the classifier. The effect of ReliefF-PCA based feature selection is analyzed for each of the seven (7)
considered chronic disease datasets.

As observed, the presented framework yields excellent results with both low and high dimensional disease datasets. As Table 2 unveils, the presented hybrid feature selection removes approximately 40% of the features from all chronic disease datasets. For ‘Colon Cancer Dataset,’ the dimensionality significantly reduces to 40.6% with the presented feature selection technique. For ‘Leukemia Cancer Dataset’, 3,394 features are selected out of 7,129 attributes. Likewise, the number of features gets significantly reduced to 1,436 from 2,880 for ‘Lung Cancer Dataset’. The performance of the hybrid model is found to be best for ‘Heart Disease Dataset,’ which removes more than 38% features from the dataset. Finally, it may be concluded from Table 2 that the proposed hybrid feature selection approach has a positive contribution on the classifier performance in Phase-II of the research.

Results of the presented framework based on ‘selected features’ with large-sized datasets are shown in Figure 3. Similarly, the analysis with small-sized datasets is depicted in Figure 4. The pseudo code of the proposed approach is shown in Figure 2:

**Figure 2. Pseudo code of the proposed approach**

```
BEGIN
Store Dataset

IF (Pre-Processing)
   Apply Pre-Processing
   IF (Column Values are Numerical)
      Remove blank/NaN values with mean of the Column
   ENDIF
   IF (Column Values are Categorical)
      Remove blank/NaN values based on mode of the Column
   ENDIF

ENDIF
Store Result in Disease Dataset

SEPARATE Class (Class) & Data (Data Without Class) from Disease Dataset
Process Data Without Class, Class with ReliefF
Store Weights obtained from ReliefF as Wm
Prepare Threshold (Theta) with Wm
Select appropriate features.
FOR Wm ≥ Theta
   Add column into new variable Selected List
ENDFOR
Apply PCA on Selected List
Store Score generated as Hybrid FS
Prepare Training (Tt) and Testing (Tt) from Hybrid FS
Make a grid space of parameters cost & gamma
Using Tt optimize cost & gamma
Choose optimized parameters and apply SVM Classification on Tt
Store Accuracy

END
```
5.2. Evaluating the Classification Accuracy of the Proposed Framework using Cross-Validation Method

The classification accuracy of the proposed model is evaluated using cross-validation method. In this work, m-fold cross-validation is applied in which the original dataset is randomly subdivided into m equal subsets. Of the m subsets, a single subset is taken as the validation data while the remaining m-1 subsets are used as training data. This process is then repeated m times, with each of the subset being used exactly once as the validation data. In this work, the classification accuracy using standard SVM classifier and the proposed system is calculated as the average of 10 observations.

Table 3 depicts the accuracy of the proposed system, which in turn shows the ability of the classifier to meaningfully classify positive v. negative classes. Moreover, it also depicts the effect of RelieF-PCA based hybrid feature selection approach on the proposed classification via optimized SVM. Accordingly, the classification accuracies are also shown when the proposed dimensionality reduction method is not applied on these datasets. The second column of Table 3 represents the classification results with ‘Standard SVM’ classifier without using any feature selection method.
The last column of Table 3 depicts the efficiency of the proposed optimized SVM classifier at the end of applying proposed ReliefF-PCA based hybrid feature selection technique. It can be seen that highest achieved accuracy is 98.5% for ‘Chronic Kidney Disease’ dataset.

As shown in Figure 6, the classification accuracy gets significantly enhanced when the proposed technique is applied on each chronic disease dataset. For low-dimensional ‘Heart Disease Dataset’ and high-dimensional ‘Colon Cancer Dataset’, the classification accuracy increases from 75.5% to 92.1% and from 66.6% to 88.5% respectively with the usage of presented method. When Table 3 is inspected, it can be observed that highest increase in accuracy comes out to be 33.6% for ‘Chronic Kidney Disease Dataset’. Although there is a slight increase in accuracy for ‘Lung Cancer Dataset’, yet it can help in the diagnosis; of course, it is better than the accuracy results with ‘Standard SVM’ classifier.

Figure 5 depicts the graphical comparison of classification accuracies of ‘Standard SVM classifier’ with the ‘Proposed System’.

Table 3. Comparison of classification accuracies of the Proposed System with Standard SVM Classifier

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Standard SVM</th>
<th>Proposed System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon Cancer</td>
<td>66.6667</td>
<td>88.5</td>
</tr>
<tr>
<td>Leukemia Cancer</td>
<td>78.5714</td>
<td>96.5</td>
</tr>
<tr>
<td>Lung Cancer (Michigan)</td>
<td>96.2963</td>
<td>96.43</td>
</tr>
<tr>
<td>Lung Cancer (Ontario)</td>
<td>77.7778</td>
<td>82.32</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>63.8655</td>
<td>98.5</td>
</tr>
<tr>
<td>Hepatitis Disease</td>
<td>84.7826</td>
<td>88.92</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>75.5556</td>
<td>92.1</td>
</tr>
</tbody>
</table>
5.3. Comparison of the Time Taken for Execution of the Proposed Framework with and without Dimensionality Reduction

This section depicts the comparison of the execution time of the proposed framework with and without dimensionality reduction. The second column of Table 4 shows the time taken for execution of the proposed system without the application of proposed feature selection technique. The third column of the Table 4 shows the time taken for execution of both Phase-I and Phase-II for all datasets.

With the resulting dimensionality reduction for all chronic disease datasets, the proposed system drastically reduces computation time. Moreover, the application of the proposed hybrid feature selection technique removes undesirable and superfluous features from the dataset, thereby shortening the time taken for overall execution of each dataset with a reduced number of features. As observed in Table 4, the time taken for execution of ‘Colon Cancer Dataset’ reduces from 14.7 sec to 11.7 sec. The same drastic reduction in execution time can best be observed in high dimensional datasets, including the ‘Leukemia Cancer Dataset’ and ‘Lung Cancer (Michigan) Dataset’. For those low dimensional ‘Chronic Kidney Disease Dataset’ and ‘Heart Disease Dataset,’ the execution time also significantly reduces by 11.7 sec and 14.8 sec respectively.

Figure 6 illustrates the graphical comparison of execution time for all large-sized chronic disease datasets. As well, Figure 7 exhibits the comparison of execution time for small-sized datasets.

5.4. Comparison of the Classification Accuracy of the Proposed Hybrid Model with Existing Classifiers

The performance of the proposed machine-learning framework is also compared with existing classifiers that are applied on chronic disease datasets. The existing classification techniques considered are ‘Logistic Regression’, ‘k-Nearest Neighbor’, Ensemble - Boosted Trees’ and ‘Support vector machines.’ According to Table 5, the classification accuracy turns out to be best with the proposed system compared to other classification systems.

Figure 8 shows a pictorial representation of the aforementioned effects in the form of bar charts.
5.5. Analysis of the Proposed Hybrid Model Using Cross-validation with Different Fold Values

The performance of the proposed hybrid model on chronic disease datasets is also analyzed using cross-validation with different fold values. In this work, m-fold cross-validation is applied in which the original dataset is randomly divided into m equal subsets. As shown in Table 6, the classification accuracy is coming out to be best with m=10 compared to other fold values. The model is checked for over-fitting while applying the optimized SVM classifier using grid search method. The model performs better on testing set than the training set.

6. DISCUSSION AND ANALYSIS OF RESULTS

6.1. Results with High-Dimensional Chronic Disease Datasets

6.1.1. Dataset 1

Colon Cancer dataset contains 2000 attributes with 62 instances. The proposed method reduces 812 attributes from the dataset and takes 11 secs for execution. As shown in Table 3, the classification accuracy with the proposed technique comes out to be 88.5%, which is the highest classification accuracy that is achieved for this dataset compared to other classifiers. In fact, the next highest classification accuracy is achieved with KNN classifier and boosted trees at 66.7%.
6.1.2. Dataset 2
Leukemia Cancer dataset contains 7,129 attributes with 72 instances. The proposed system achieves 96.5% accuracy with this dataset. Without feature selection, this dataset executes in 43 sec. With the use of ReliefF-PCA based feature selection via optimized SVM classifier, the execution time decreases to 13 sec. Furthermore, the number of parameters reduces from 7,129 to 3,394 for this dataset.

6.1.3. Dataset 3
Lung Cancer (Ontario) dataset contains 2,880 attributes with 39 samples and two classes. The presented two-phase approach reduces 49.8% features from the dataset. The computational time of the model without feature selection approach is 11.2 sec. With the execution of Phase-I & Phase-II together, the time reduces to 9.2 sec. This reflects the positive effect of proposed feature selection technique on the classifier performance. Also, with the proposed approach, the classification accuracy obtained is 82.32%, which is very high compared to other classifiers (see Table 5).
6.1.4. Dataset 4

Lung Cancer (Michigan) dataset consists of 96 instances and 7,129 attributes. The presented method eradicates approximately 41% features and takes only 15 sec for execution. Compared to other existing classifiers, the proposed classification system achieves an excellent accuracy of 96.4%.

6.2. Results with Low-Dimensional Chronic Disease Datasets

6.2.1. Dataset 5

Chronic Kidney Disease dataset consists of 25 attributes and 400 samples. The algorithm performs extremely well in terms of feature selection by diminishing the dimensionality of dataset to 10 parameters. Moreover, with the execution of both phases of the research, the execution time decreases from 20.3 sec to 8.6 sec. With this dataset, the proposed system performs outstandingly well and achieves highest classification of 98.5% in comparison to other datasets.
6.2.2. Dataset 6
Hepatitis Disease dataset contains 155 samples with 19 features and two (2) classes. The proposed approach reduces dimensionality of the dataset to 12 features. Compared to existing classifiers, the proposed system performs very well with 88.92% classification accuracy. Moreover, it executes fast, that is, in 1.9 sec.

6.2.3. Dataset 7
Heart Disease dataset contains 270 samples and 13 features. For this dataset, the dimensionality of the features reduces from 13 to 5 attributes, which is approximately 38% features of the dataset. The proposed approach executes both phases in 3.6 sec and obtains 92.1% accuracy with this dataset.

6.3. Statistical Significance Test
To investigate the effectiveness of the proposed method, ‘Friedman-Test’ is applied that efficiently tests the null hypothesis of identical populations (Friedman, 1940). This test can be used for comparing the rank of q-algorithms over d-datasets.

\[
F_k = \frac{12}{dq(q+1)} \sum R^2 - 3q(q+1)
\]  

(1)

Friedman test starts with first stating the hypothesis and the significance level. It ranks all the models as mentioned in Table 7 and depending on the test statistics and calculations, it determines the value of \( F_k \) from equation (1). Next, from the critical value of the chi-squared table, null hypothesis is either accepted or rejected. Here, null and alternative hypothesis are defined as:

H₀: No difference between all five models
H₁: Difference between all five models

Here, putting values of \( d=6, q=5, R (15, 13, 13, 19, 30) \) in equation (1), \( F_k \) is achieved as 13.62. From chi-squared table for the value of \( q \) and degree of freedom (0.05), the critical value is 9.48773. Since, \( F_k \) is higher than the critical value, null hypothesis is rejected; hence, there exists statistically significant difference between the three models.

Table 7. Ranks for all comparison models

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Logistic Regression</th>
<th>K-Nearest Neighbor</th>
<th>Ensemble -Boosted Trees</th>
<th>Support Vector Machines</th>
<th>Proposed System (Phase I + Phase II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon Cancer</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Leukemia Cancer</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Lung Cancer (Michigan)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Lung Cancer (Ontario)</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Hepatitis Disease</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>13</td>
<td>13</td>
<td>19</td>
<td>30</td>
</tr>
</tbody>
</table>
7. CONCLUSION

This paper presents an efficient two-phase hybrid classification framework for the diagnosis of chronic diseases. The diagnosis is based on dimensionality reduction via a combined ReliefF-PCA and optimization of SVM classifier. The key aspect is the selection of appropriate threshold for choosing relevant features from the dataset. The two-fold contributions of this work may be stated as follows: (a) to attain reasonably good classification accuracy with possible speed up computation; and (b) to reduce high data dimensionality via proposed feature selection technique.

Experimental findings demonstrate the superiority of the presented framework compared to existing classifiers in respect of high classification accuracy. With repeated experiments and computations, it is also shown that the proposed method works well with both text and micro-array datasets in large dimensions as well as low-dimensions for classification. The proposed hybridized model gives outstanding results and significantly enhances the accuracy of classification algorithm for all chronic disease datasets.

One implication of the finding is that the presented model is a promising tool for the diagnosis of chronic disease patients. More importantly, it can assist physicians and other healthcare professionals such as health informaticians in the analysis of diagnostic results of chronic disease datasets and hence can act as a valuable adjunct tool in changing current clinical practices.

Another important implication with the presented approach is that it eliminates extraneous and surplus features with substantial reduction in the execution time for mining all datasets, thereby achieving outstanding classification accuracy. The experimental findings prove that the proposed approach outperforms other traditional classification algorithms. The application of dimensionality reduction prior to classification saves time, effort and cost of unnecessary tests, and hence, substantially reducing the diagnostic complexity and resulting computational time.

In future, the proposed methodology can be modified to provide diagnostic results in minimal time. The model can be extended with parallelism through GPUs that can more efficiently execute high-dimensional datasets. Secondly, the proposed methodology can also be applied for assessing other high-dimensional chronic disease datasets.
REFERENCES


