The life cycle of the developed research work is initially implemented by modifying the source code of Weka 3.5.8 (http://www.cs.waikato.ac.nz/~ml/weka/). The modifications started with creating a method repository by collecting the available methods and functions of attribute selection algorithms. The attribute selection algorithms are grouped into three distinct classes based on ranking, correlation and clustering. In the next phase, changes are made so that any dataset that is passed to the prototype undergoes attribute subset parallel on all the three distinct classes and the output is noted for all the algorithms available in the repository. Initially a small dataset with few numbers of attribute is tested in different formats and then it is tested for large number of attributes.

An adaptive learning algorithm is then developed and tested for the functionality and integrated with Weka. The aggregation part is then developed and integrated with Weka and a normalized subset of attributes is obtained after aggregating information gain, correlation and clustering outputs. The prototype is further extended in incorporating missing values by modifying K-Means algorithm and then extended to C4.5. The training and classification part available in the Weka is used...
for initial simulation purpose. The classification accuracy is also initially tested using the available code from Weka.

The prototype is then implemented using Java and tested for small datasets by incorporating some of the available classes from Weka for any type of datasets. A part of the simulation is also carried out using PolyAnalyst 5.0 data mining suite and tested the results with the prototype that is implemented.

The prototype is tested and the results are correlated for seven Datasets from UCI Machine Learning Repository ranging from 10 attributes to 10000 attributes (http://archive.ics.uci.edu/ml/datasets.html). The datasets used are Poker (10 Attributes), Adult (15 Attributes), Covertype (54 Attributes), Madelon (500 Attributes), Dorothea (770 Attributes), Gisette (5000 Attributes) and Arcene (10000 Attributes). The results are tabulated and represented graphically. The adult dataset from UCI machine learning repository is clearly examined for different output to check that this approach works normal in giving the deliverables that other attribute selection technique gives.

A comprehensive study of the selected attributes using the developed method is conducted. The performance of the developed method is analyzed using domain knowledge for validation. The initial step is to
find the number of attributes selected by the various methods. This is dependent of course on the threshold parameters. In this research work, the threshold is calculated by the adaptive aggregation algorithm which selects the separation measure. It is found that this hybrid approach gives more information (Figure 5A) on the attributes that are missed in attribute selection technique based on correlation. The approach works equally well for datasets with large number of attributes. Some of the results obtained by this approach for attributes less than 100 and greater than 100 are tabulated and represented graphically.
It is also found that even when the number of attributes is high the aggregation of attribute selection works (Figure 5B) better in giving the number of necessary and important attributes because it takes account on correlation, ranking and distance based clustering technique. It is also observed that there are variations in the distribution characteristic
in the methods like information gain, clustering and separability in different datasets whereas the developed method works uniformly irrespective of the different datasets. This proves that the developed approach is domain independent and works uniformly for datasets irrespective of the number of attributes.

Dataset Vs Number of Attributes

![Dataset Vs Number of Attributes](image)

**Figure 5b:** Dataset with number of attributes greater than 100

The number of attributes obtained for the seven datasets on different methods for specific output fields is given in table 5A. The developed hybrid method is also highlighted.
In Table 5A, the Information Gain for the dataset is calculated using the formula $IG = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y)$. The parameters involved in this formula are explained in page 49. The aggregation of the dataset is calculated using the pseudocode given in page 84.

When the developed hybrid algorithm works for few number of attributes dataset, the attribute subset obtained from separability measure and cluster relationship is similar in number. Whereas, when the numbers of the attributes are more, significant difference is encountered; this validates that the developed hybrid algorithm brings out relevant attributes that are apart from the methods given by pairwise and separability measure for a particular output, thus increasing the clarity of the mining result. The results listed in table 5A also give an insight that the subset of attributes resulted from information gain and cluster relationship is relatively high since information gain ranks the attributes on certain factors, even after giving threshold condition; the same is the case in cluster relationship since the class distance is taken into measure even after applying threshold condition, the results are high when compared with pairwise and separability measure. The results

### Table 5A. Dataset with number of attributes

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No of Attributes</th>
<th>Pairwise</th>
<th>Information Gain</th>
<th>Separability</th>
<th>Cluster Relationship</th>
<th>Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poker</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Adult</td>
<td>15</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Covertype</td>
<td>54</td>
<td>11</td>
<td>42</td>
<td>18</td>
<td>47</td>
<td>21</td>
</tr>
<tr>
<td>Madelon</td>
<td>500</td>
<td>9</td>
<td>208</td>
<td>11</td>
<td>103</td>
<td>12</td>
</tr>
<tr>
<td>Dorothea</td>
<td>770</td>
<td>205</td>
<td>282</td>
<td>461</td>
<td>638</td>
<td>298</td>
</tr>
<tr>
<td>Gisette</td>
<td>5000</td>
<td>1253</td>
<td>2387</td>
<td>1191</td>
<td>1405</td>
<td>1338</td>
</tr>
<tr>
<td>Arcene</td>
<td>10000</td>
<td>689</td>
<td>3791</td>
<td>832</td>
<td>273</td>
<td>456</td>
</tr>
</tbody>
</table>
obtained for Gisette dataset seems to be abnormal for the information gain methods because of the nature of the dataset. The reason for Gisette dataset is the selection of the target attribute in the information gain methods that relatively gave the subset with more attributes when compared with other methods used in this work. The problem of this dataset is to classify the highly confusable digits in the handwritten digit recognition.

The developed approach is applied on datasets with different number of attributes and characteristics and the resultant attribute subsets are checked for accuracy. The classification accuracy obtained using neural network classifier for aggregation of attributes is given in figure 5c. This shows that the attribute subset selected by the developed approach is well within the prescribed criteria of accuracy given in literature [Legrand and Nicoloyannis, 2005; She et al., 2005; Chen et al., 2008]. This validates the conclusions on the aggregation based hybrid algorithm.

![Classification Accuracy - Aggregation of attributes](image)

*Figure 5c: Classification accuracy based on aggregation of attributes*
A comparative study for the classification accuracy on different datasets on different attribute selection methods is carried out with the same classifier. The results are shown in figure 5D, which highlights that in most of the cases, the hybrid algorithm that is developed outperforms other methods discussed in the literature.

**Classification Accuracy of Different Data Sets With Different Attribute Selection Methods**

![Classification Accuracy Graph](image)

**Dataset**
- Adult
- Dorothea Covertype
- Gisette
- Madelon
- Poker
- Arcene

- Pairwise
- Information Gain
- Separability
- Cluster
- Aggregation

*Figure 4D: Comparison based on percentage with different attribute selection Methods*
Figure 5E highlights the comparison on classification accuracy for different datasets on a different view.

![Comparison Based on Percentage](Diagram)

*Figure 5E: Comparison based on percentage for different datasets*

The observation made from the figure 5d and 5e states that, for Gisette dataset the percentage for almost all the methods lies between 52 to 62 percentages. Thus this type of dataset has to be given more importance on other methods so as to get more information from such datasets.

Comparisons are also carried out after imputing the missing values using C4.5 and K-Means methods for the attributes obtained by hybrid algorithm. The results are shown in figure 5F.
The result clearly shows that there is a significant increase in classification accuracy when the missing values are imputed using C4.5 and K-Means method; also in most of the cases imputation carried out using K-Means algorithm outperforms C4.5 methods. This also gives an insight that K-Means algorithm can be used to impute the missing values for better accuracy.
The classification accuracy for all the methods used in this research work as well as the hybrid method including the imputation of missing value using C4.5 and K-Means is shown in figure 5g. The last three results shown in the figure are the results highlighted by this research work. The other results are taken based on the literatures that are reviewed. The target variable used may vary from method to methods but the results are computed based on the algorithm proposed in literature.

![Classification Accuracy of Different Data Sets With Different Attribute Selection Methods](image)

Figure 5g. Classification accuracy after incorporating missing values for the hybrid method as well as with other attributes selection methods

Thus the hybrid algorithm that is developed in this research work gives a better attribute selection algorithm by aggregating the attributes subsets obtained from information gain, correlation and clustering methods. Some of the simulated outputs obtained during the process of this research work are also given below. Figure 5h shows the pairwise
correlation of attributes for a dataset whose output attribute is given in x axis and range of other attributes are given in y axis.

The above figure provides information that for the particular output attribute, other set of attribute correlates i.e. most of the attributes makes almost the same path from location 0 to 140. Thus the pairwise correlation measures are viewed for the set of attributes obtained by the hybrid method.

*Figure 5H: Pairwise correlation measures*
The separability measure for the set of attributes is shown in the figure 5i.

The output attribute is considered along x-axis and a subset of an attribute is considered along y-axis. It is found that the density is very high during the initial stages and low as the value on x and y increases.

Figure 5i: Separability measure

The output attribute is considered along x-axis and a subset of an attribute is considered along y-axis. It is found that the density is very high during the initial stages and low as the value on x and y increases.
The snapshot of the likelihood measure using the likelihood ratio is shown in figure 5j.