CHAPTER 3
HADOOP-BASED ENSEMBLE CLASSIFICATION MODEL FOR DISTRIBUTED DATABASES

There is a super abundance of data from myriad sources as they emanate from sundry fields called distributed databases. They happen to be manifold, intricate, dissimilar and autonomous. Because of the substantial number of occurrences with excessive and extraneous features, a suitable and fitting filtering model has been employed to fill the noisy features or occurrences. In order to deal with a huge number of characteristics, an effectual attribute assortment model was employed to to select the ranked features for classification or clustering. Al-Khateeb and Masud [79] applied a succinct set of rules utilizing feature selection model in rough set theory. However, a vital challenge in the classification algorithms is the error rate along with class imbalance.

Generally, high-dimensional document spaces are hard to handle, pre-process and classify due to large amounts of document sets. To improve the learning of the classification algorithm, a large number of samples are required to be learnt according to their dimensions. Conceptually, this document space is a sub-space of low dimensionality and it is wrapped with ambient space. Due to this dimensionality issue, many dimension reduction methods were developed to resolve this issue. The key purpose of this method is to decrease the document dimensions and enhance the performance and efficiency. The fundamental difficulties in dealing with online medical databases are arithmetic computing procedures and data accessing, semantics and domain knowledge with several high-volume data applications. In addition, there are several challenges arising because they are complicated, continually fluctuating besides noise in the growing data.

3.1 Document Clustering

These days, with the fast-paced development of internet facilities and web know-hows, document data is accessed, stored and shared over the web with considerable
ease. One of the challenges for the researchers has been how to facilitate organizing all these huge clusters of documents so as enable users to search and find information they are looking for quickly and efficaciously, Document Clustering (DC) is one of the most important methods employed in organizing documents in a process that is without intervention or supervision. Because of its capability of being applied in several domains, people are showing a lot of interest in this area called document clustering.

3.1.1 Non-negative Matrix Factorization (NMF)

The fundamental principle of NMF technique is clustering the non-negative features of biomedical documents. NMF can be categorized under the unsupervised learning scheme and is responsible for retrieving relational features of the high dimensional data. This algorithm is based on linear segmentation technique which emphasizes on non-negative data. The primary objective of this approach is to gather alternative inherent structures among data and reduce the dimension and convert document term matrix into smaller ones. These smaller matrices must contain elements with non-negative values. There are many applications of NMF algorithm such as pattern recognition, text clustering and bioinformatics.

3.1.2 Support Vector Data Description (SVDD)

SVDD is one of the most used clustering algorithms which follow the basic principles of SVM classification scheme. It performs the mapping of data objects to high dimensional feature space. SVDD uses non-linear transformation function for mapping [80]. It also deals with the detection of outliers that have almost all the data points in a specific feature space. A boundary is formed around the outliers by combining the closed contours.

These contours are named as cluster boundaries and all the points of an individual contour are included under the same cluster. The proposed SVDD algorithm outperforms the conventional algorithms by creating cluster boundaries of any shape and size irrespective of the available data side. Traditional techniques have a number of issues, which can be overcome by SVDD clustering approach. The proper implementation of SVDD in the process of document clustering is still an open
research problem. There exists no optimized scheme for document clustering, which resolves all issues of traditional approaches and provides better performance.

3.2 Hadoop based Data partition

The Map-Reduce structure does a variety of jobs like underpinning and supporting large database with high feature space, partitioning them into slighter sets and dispensing them to diverse cloud groups for calculation. Map-Reduce outlooks data as a key, value pair as <Key, Value> as shown in Figure 3.1. Data-intensive processing is currently well thoroughly considered with focus around the Map-Reduce structure for vast scale data investigation. Progressively, it is a fault-tolerant and scalable information handling device, which conveys the capability to figure and process substantial volumes of datasets nearby a few low-end computing cluster nodes. Be that as it may, Map-Reduce have natural difficulties in its proficiency and execution. In this manner, many research works have been devoted to addressing the difficulties of the Map-Reduce structure [81].

![Figure 3.1 General Workflow of Map-Reduce Framework](image)

The primary target of knowledge mining strategies is to discover useful info from huge informational indexes called datasets. The knowledge discovered in this process
helps a lot in decision-making frameworks. However, if the intensity increases, the size and dimension of the data set size hurt the computational competence as well as the time taken to process the same. However, many of the existing models designed for optimization are not good at parallel computations and the parallelization of the traditional algorithms are difficult to put into practice as they are non-trivial.

The decision tree is the key classification strategy in data mining. These days, analysts of decision tree algorithm concentrate on improving the efficacy of the data processing competence. As the improvement of the systems networking and real-time applications upsurge, the dimensions of information likewise increment proportionately. With a specific end goal to determine these issues, a parallel distributed decision tree structure utilizing the Hadoop framework system is utilized to address these voluminous clusters of information.

3.3 Biomedical XML document preprocessing

The most widely used applications of text mining hail from biomedical domain due to a large number of medical document sets. In the traditional text mining approaches, the search process depends only on "sorting and motif" in the case of open access articles. Biomedical open access articles are extracted in full text and licensed under the creative, common-license. Around 100GB of articles are detected in the first step of the document extraction algorithm. Unambiguous structured data are essential for the better performance of extraction model. Besides XML tags, generally, all the documents from the PubMed repository have unstructured information. The document clustering approach is responsible for pre-processing and clustering full-text articles. Unlike other text mining approaches in the biomedical domain, this approach emphasizes on abstracts.

Of late, XML is used as a typical format for information sharing on the web. The most common and frequently implemented approach is clustering. Here, clustering represents merging of similar types of XML data and applications of XML clustering. They are retrieval of information, integration of data, classification of documents, web mining as well as processing of queries.
The major issues in XML data preprocessing for clustering are described below [82]:

- Initially, the clustering process calculates the similarity index among numbers of different XML data. Nevertheless, evaluating the similarity function is a major problem because of the heterogeneity property of XML documents.

- Implicit dimensionality has increased to a great extent.

Biomedical documents, phrases, sentences are used in the feature extraction to extract the main features of the original documents. The graph-based feature extraction generates the feature extraction by mining phrases or sentences from the set of key peer nodes of the overlay network. Finally, key phrases or sentences are extracted by computing the ranking scores and then selecting the highest scored phrases or sentences.

**Input MEDLINE XML file:**

```xml
<?xml

<MEDLINE Citation>

<PMID>267801</PMID>

<Article Title>

Hyperostosis on the alveolar process margin

</Article Title>

</MEDLINE Citation>

<MEDLINE Citation>

<PMID>267802</PMID>

<ArticleTitle>

The role of supplementary canaliculi in the etiology and therapy of periapical lesions

</ArticleTitle>

</MEDLINECitation>

<MEDLINECitation>
```
Prosthetic rehabilitation by means of removable dentures with wires for the distribution of forces in Kennedy class 2a and 2a modification 1

The typical use of Boutycin (indomethacin) in odontostomatology, Clinical study

The effects of indomethacin applied on the surface for postoperative pain and oedema have been studied. The effect is positive and is predominantly obvious in the case of pain.

Physician manpower in Missouri: 1975.
<ArticleTitle>
Treatment of tension and migraine headaches with biofeedback techniques
</ArticleTitle>
</MEDLINECitation>
</xml>

3.4 Traditional feature selection models on biomedical databases

A novel Bayesian Network (BN) approach [83] is developed to classify the documents using MeSH terms. This model performs classification using the entity association of different MeSH terms and can easily represent the conditional independences among MeSH terms and MeSH ontology. This technique is implemented to represent the resources of MEDLINE documents with distinct abstraction levels. To solve these issues, medical ontologies are used to find the MeSH synonyms for medical concepts. Machine learning algorithms are generally implemented in the process of textual classification scheme through an inductive approach.

Additionally, an extension of the Bayesian network is carried out to classify MEDLINE documents using SVM based BN classifier. Both BN and SVM based BN classifiers are sensitive to balancing strategy performance analysis. Class-imbalance problem is common in the area of text categorization and classification models. These approaches are not able to solve the traditional class-imbalanced problem. So, many advanced approaches are developed to improve the accuracy rate of standard classification techniques. These generalized classification approaches involve the following schemes: sampling approaches, cost-sensitive techniques, recognition-based methods and active learning techniques. Sampling approaches have been used to solve the class imbalance problem by removing randomized data from majority class. These techniques are known as under-sampling techniques. Sometimes, few additional randomized data are included in the minority class and this phenomenon is called as over-sampling [84]. Cost-sensitive learning method uses cost-matrix for classifying
the document categories on small instances. Recognition-based methods learn rules from the minority class without using the instances of the majority class. Active learning methods are employed to resolve the complications related to unlabeled data.

Feature selection based learning techniques are modelled to find the hidden learning rules in the minority class instances, which may or may not use instances of the majority class. Bayesian-based feature selection models act as a probabilistic framework to classify MEDLINE documents automatically. Bayesian models are not applicable when a document is labelled with a specific MeSH term or if it is also linked to the ancestors of the term in the hierarchical model.

Let us consider a case in point; a document is indexed with terms A01.047.025.600.451 can be symbolized by the terms A01, A01.047, A01.047.025 and A01.047.025.600 correspondingly. This supposition enhances ontology-based document feature selection technique on MEDLINE documents.

The random sampling technique is proposed in [85] is to remove the randomly selected elements from the over-sized majority class. This process continues until it matches the size of the minority class and cost-sensitive learning scheme. It includes the modifications of the relative cost associated with misclassification of positive and negative class. To apply machine learning approaches for feature extraction, each individual document is required to be transformed into a feature vector. Several feature selection approaches have been implemented in order to resolve the issue created due to the high-dimensionality of features space. Many statistical and probabilistic machine learning methods have been implemented in the process of document feature selection for years. It includes K-nearest neighbour approach, Bayesian technique, decision tree scheme, symbolic rule learning method and neural networks approach.

Feature learning models are based on scalable Machine Learning approaches and Rule-based reasoning methods. This process is categorized under unsupervised learning because there is no previous classified example or constraint related to data items having labels. The classification model is presented by ontology (Abox+Tbox). It is automatically learned out of a large amount of data by scalable Machine Learning
and Big Data techniques. The process of feature extraction has a high significance in the field of classification. All features are divided into two groups, they are:

- Feature extraction using noisy attributes and contextual information is performed on the limited document sets.
- Extracts correlated features.

Traditional feature extraction models discard noisy features to decrease the high dimensional features to a lower dimensional feature. Many approaches are developed in the field of medical disease prediction such as cancer prediction, gene prediction and protein detection etc. In [86], the rule mining approach is integrated with principal component analysis (PCA) and back propagation to predict and detect the chances of microarray cancer. A modular neural network is developed to recognize and analyze the cardiac diseases using Gravitational Search Algorithm (GSA) and fuzzy logic algorithm for better performance [87].

**3.5 Traditional Document Classification models**

A substantial number of machine learning approaches have been produced to separate significant and fascinating features all the more proficiently on little datasets [88]. As the size of the biomedical repositories increases, an individual machine is not capable of handling high-dimensional features due to limited computing resources. Therefore, many research efforts have been carried out on parallel and distributed computing. Machine learning approaches are implemented to execute large-scale models in parallel and distributed environments. The Map-Reduce framework is used to develop parallel and distributed models on high dimensional data. An abstraction-based model is developed on a single hardware interface which permits the programmer to integrate their models without any computing issues. Multiple-instance learning [89] is an efficient approach in the field of bioinformatics and machine learning. In this learning approach, the training dataset consists of a series of positive feature space B+ and negative feature space B−. A feature space is considered to be positive, if and only if it has at least a single positive instance. At rest all cases, the feature space is considered to be negative. A group of positive feature
space is merged to form a positive class, whereas many of the negative feature spaces are integrated to form negative class. The main objective of multiple-instance learning approach is to train classifiers with the help of training data set and then to predict the labels of unknown feature space.

Hierarchical Multi-Label Classification (HMC) technique can be defined as a conventional decision tree approach which merges multiple instance classes with the hierarchical manner. In this technique, biomedical documents are correlated with the hierarchical paths to classify the multi-class labels in the decision tree. This technique is used to classify a large number of unstructured text documents as it is impossible for the single system to adopt certain modifications.

Most of the traditional multi-instance based decision tree models are broadly categorized into two types: Bag-level techniques and Instance-level techniques. In bag-level techniques, the total bag is assumed to be a training unit and the bag-level classifier is used to predict the bag label. Diverse Density Support Vector Machines (DD-SVM) is a bag-level based approach designed to integrate the feature classes using the Diverse Density (DD) function and nonlinear mapping. Both of these measures are used to map each bag to a point through instance feature classes. The instance label techniques are Multi-class SVM [90] techniques based on the traditional SVM scheme to classify large number instances in the training bags. Each positive bag contains at least a single instance that is classified as positive through the classification algorithm. Traditionally, two types of classification models are used to classify the biomedical data streams, such models are incremental learning and chunk-based learning. Incremental learning is used to classify or predict the test document data using historical streaming data so it is able to adapt the changing of concepts. On the contrary, the Chunk-based learning scheme decomposed the whole data flow into a series of data lumps. This technique integrates the base classifiers to form an ensemble classifier for forecast. Several ensemble learning approaches are developed to improve the true positive rate of the base classifiers for ensemble classifier. These models include weighted voting, dynamic voting, majority voting, Adaboost and mean weighted selection. Using the base classifiers and data size, the ensemble
learning model can achieve better prediction accuracy as compared to a weak individual classifier.

The feed-forward neural network can be considered to be the most commonly and widely implemented neural network. This method has one or more hidden layer(s) along with an output layer. The output layer is responsible for transmitting a final response on the training dataset. A large number of research works have been implemented in the field of Feed-Forward Neural Networks since years. This model has linear or nonlinear structure directly mapping from inputs. These structures are not appropriate for classical parametric constraints to manage large inputs in the traditional models. Another important feature of the Feed-Forward Neural Network is the inter dependency among the layers through parameter mapping. Single-Hidden-Layer Feed-Forward Networks (SLFNs) are treated as the most efficient and widely used Feed-Forward Neural Networks on small datasets.

**Weighted Extreme Learning Machine Technique (Weighted ELM)**

In order to resolve the issues of traditional 'Extreme Learning Machine' (ELM), the weighted-ELM approach got developed subsequently. In case of large sample size, the weights are increased gradually on time. In most of the feed-forward ANN techniques, parameters of each layer are required to be tuned through several learning approaches. Gradient Descent-Based Approaches and Back-Propagation (BP) techniques are some important learning algorithms in Feed-Forward Neural Networks [91]. The speed of learning models is very slow in case of Feed-Forward Neural Networks compared to ANN because of better generalization capability and fast computational speed. A lot of problems are detected in the case of conventional Gradient-Descent Algorithms like stopping criterion, learning rate, a number of epochs and local minima.

These approaches are very slow because of improper learning phases or coverage of local minimum. Additionally, a lot of iterative learning schemes are developed in order to achieve better learning efficiency and performance. This approach includes neurons in an increasing order and completely depends upon kernel-based approaches. In the case of Bayesian ELM, the hidden layer is not tuned unlike training of conventional Feed-Forward Network. All the parameters like input weights, output
weights and biases of hidden layers are selected randomly to minimize the training error. The major disadvantages of the above schemes are slower learning rate of networks and reduced capability to manage the complexities.

Fuzzy ELM (F-ELM) approach involves two basic phases known as training and prediction. In the training phase, a three-layered architecture is proposed. Similar to ELM method, the training phase iteratively generates scaling parameters in concealed layer and weights in the concealed layer and output layer. Hence, the functions of F-ELM are similar to the fuzzy inference system. In the prediction phase, an input feature vector is mapped into a trained F-ELM algorithm which evaluates the final outputs.

Decision trees are represented as directed graphs having nodes and edges. The root node and the intermediate nodes always represent tests, whereas the conditions on nodes are represented as outgoing edges. Besides these, the leaf nodes represent different class labels which are responsible for determining the hidden patterns in large databases. These are responsible for predicting the decision patterns in the decision tree. With the increase of k-tree size, the amount of information loss also increases in the decision tree construction. These models perform a corresponding hierarchical based selection of features algorithm to arrive at the decision rules for every level. The major drawbacks of this attribute selection technique are: 1) as the noise of attributes in single and multi leveled space increases, time and space for computation additionally increases. 2) This model relies upon entropy measures on the information distribution uniformity.

![Figure 3.2 Conventional (a) C4.5 and (b) ELM decision Tree classifier](image)
ELM based decision tree [92] is used to find the infrequent decision rules using logistic regression function i.e., a new occurrence is arranged by navigating the tree from the root node to the leaf node and decision on the attribute is made on the class attribute value of the leaf node using regression analysis as shown in Figure 3.2.

The ELM decision tree is not applicable to incremental tree formation on high dimensional databases. Additionally, this approach doesn't support mixed attribute types utilizing Map-Reduce framework.

Limitations

- The performance of the tree depends on the number of hadoop nodes and the memory constraints.
- It is essential to fix the static constraints of the minimum class instances on the intermediate mappers to get better performance.
- When a Map-Reduce framework is under the training phase, it is essential to develop the static classification and static parameters

3.6 Feature Selection and Decision-Making using Hadoop framework

To finding the most pertinent feature selection measure from a given data set with discretized attribute values, the rough set technique is designed and implemented in [93]. Lower and Upper approximations of decisive \( D_c \) with respect to a partition \( \pi_{Att} \) are defined as:

Lower bound approximation \( App{x}^l_A(D_c) = \{k \in U / [k]_A \subseteq D_c \} \);

Upper bound Approximation \( App{x}^u_A(D_c) = \{k \in U / [k]_A \subseteq D_c \neq NULL \} \);

An entropy measure used to reduce the features size is given as

\[
Info(D) = - \sum_{p=0}^{N} (D^p / D)^* \log(D^p / D)
\]

\[
Info(D / A_i) = - \sum_{p=1}^{\mid D \mid} (D^p / D) \sum_{p=0}^{\mid D \mid} (D^i_p / D^p) \ast \log(D^i_p / D^p)
\]
Where \( D_p \), \( D_i \) and \( D^i \) denotes the number of the objects, the number of objects equal to \( i \) on \( D \) in Attribute \( r \) and the number of objects equal to \( I \) on \( D \) respectively.

The process of feature extraction has high significance in the field of classification. All features are divided into two groups, those are:

1) According to the first group, features extraction using noisy attributes and contextual information.

2) The second group contains correlated features.

Traditional feature extraction models discard noisy features in order to decrease the high dimensional features to a lower dimensional feature.

Let us assume \( \text{Nmin} \) and \( \text{Nmax} \) are minimum and maximum numbers of hidden neurons respectively, where \( N \) denotes the present value of hidden neurons. For each and every \( N \), the average accuracy rate of EL through 10-fold cross-validation scheme is evaluated. At last, hidden neurons having maximum average accuracy is chosen as optimal. After selecting the optimal numbers of hidden neurons, the EL classifier is implemented in order to evaluate the classification accuracy by considering the outcomes of PCA and the outcomes are averaged later.

The training datasets used in this paper have a significant issue for any classification models as they have large number of feature space, ranging from 100 to 12000 features. The larger the feature space increases the search space and computational memory for disease prediction. Another crucial issue for handling the high dimensional features is the small sample size problem. The accuracy of the model employed will be reduced if the size of the training data is not sufficient relative to the feature space.

In the past, machine learning models used a single classification model to forecast the test dataset using the training instances. However, multiple classifiers can be used to forecast the same test dataset using the training instances. This process is known as ensemble learning. Ensemble classification has been modeled to different imbalanced problems to improve the classification accuracy using the optimal feature selection measures.
Particle swarm optimization (PSO) is very popular optimization techniques in machine learning models. PSO is generally applied in the literature to adjust the initialization parameters of base classifiers in the ensemble learning models. The main idea of this model is to optimize the traditional PSO parameters in the ensemble classification model in order to improve the accuracy and error rate. In the ensemble model, neural network is used as one of the base classifier and weights are initialized using the proposed PSO technique.

In this proposed rough-set-based decision tree model as shown in Figure 3.3, decision patterns are constructed at the individual level in the multi-dimensional model. The main enhancement of this technique is to filter the decision patterns with insignificant rule set. Finding and extracting rules on each decision tree level depend on the multi-dimensional information show.

![Figure 3.3 Random Forest Using MapReduce](image)

This system has two stages: one is data Reduction stage and the other is decision rule generation stages. In the primary stage, attribute reduction and attribute filtering measures are connected on the input data set. Also, in the second stage, decision rules
utilizing hierarchical rough-set and decipher the connection between the decisions rules separated in each level of the hierarchy.

3.7 Experimental Results

**Amazon Network Setup:** Let us consider an Amazon web services EC2, it has three different types of storage sizes such as small, medium, large and this type including various sorts of cloud resources. These instances and services are available for different country timings and zones. Table 3.1, illustrates the Asia Instance servers and their pricelist of the micro to large instance types for the Windows Usage machines and LINUX/UNIX [95].

Amazon Elastic Compute Cloud (Amazon EC2) provides flexible-sized computation capability of the cloud. EC2 provides simple and easy web-scale computation for developers. A user-friendly interface is available for configuration of hardware capacity needed to the finest detail, with a very little amount of effort. The above interface is responsible for inclusion and deletion of instances. These EC2 cloud services are placed in either private or public Cloud storage severs by permitting the customer to determine which instances should be disclosed to the internet. In order to implement Amazon EC2 instance, at first, a pre-configured AMI, network rights and security and the uploading of a developed web application are required. The users are required to choose the starting and ending of instance according to their computing needs. A number of instances must have the number of AMI. The activities of newly created instances are monitored in the next step to configure other application services. In order to handle both inbound and outbound network access, security groups and network Access Control List (ACLs) are implemented.
Table 3.1 Amazon EC2 Setup Configurations

<table>
<thead>
<tr>
<th></th>
<th>Linux/UNIX Usage</th>
<th>Windows Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Purpose - Current Generation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m3.medium</td>
<td>$0.0086 /Hour</td>
<td>$0.0591 /Hour</td>
</tr>
<tr>
<td>m3.large</td>
<td>$0.0162 /Hour</td>
<td>$0.1171 /Hour</td>
</tr>
<tr>
<td>m3.xlarge</td>
<td>$0.036 /Hour</td>
<td>$0.2201 /Hour</td>
</tr>
<tr>
<td>m3.2xlarge</td>
<td>$0.1539 /Hour</td>
<td>$0.4391 /Hour</td>
</tr>
<tr>
<td><strong>Memory Optimized - Current Generation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r3.large</td>
<td>$0.0175 /Hour</td>
<td>$0.1741 /Hour</td>
</tr>
<tr>
<td>r3.xlarge</td>
<td>$0.0478 /Hour</td>
<td>$0.2821 /Hour</td>
</tr>
<tr>
<td>r3.2xlarge</td>
<td>$0.1785 /Hour</td>
<td>$0.4442 /Hour</td>
</tr>
<tr>
<td>r3.4xlarge</td>
<td>$0.48 /Hour</td>
<td>$0.6744 /Hour</td>
</tr>
<tr>
<td>r3.8xlarge</td>
<td>$0.8791 /Hour</td>
<td>$0.9583 /Hour</td>
</tr>
<tr>
<td><strong>General Purpose - Previous Generation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m1.small</td>
<td>$0.0081 /Hour</td>
<td>$0.0261 /Hour</td>
</tr>
<tr>
<td>m1.medium</td>
<td>$0.0102 /Hour</td>
<td>$0.0531 /Hour</td>
</tr>
<tr>
<td>m1.large</td>
<td>$0.0175 /Hour</td>
<td>$0.1061 /Hour</td>
</tr>
<tr>
<td>m1.xlarge</td>
<td>$0.034 /Hour</td>
<td>$0.2111 /Hour</td>
</tr>
<tr>
<td><strong>Memory Optimized - Previous Generation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m2.xlarge</td>
<td>$0.0166 /Hour</td>
<td>$0.1121 /Hour</td>
</tr>
<tr>
<td>m2.2xlarge</td>
<td>$0.0378 /Hour</td>
<td>$0.2241 /Hour</td>
</tr>
</tbody>
</table>
Experimental results are performed on different biomedical documents in XML format. Each data set has different types of document types in MEDLINE Repository.

Table 3.2 Classification models for feature selection using Rough-set

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mapper Time (mins)</th>
<th>Reducer Time (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA_FSA</td>
<td>15.35</td>
<td>4.57</td>
</tr>
<tr>
<td>Neural Network</td>
<td>21.53</td>
<td>5.75</td>
</tr>
<tr>
<td>ELM-Tree</td>
<td>14.65</td>
<td>5.36</td>
</tr>
<tr>
<td>Random Forest (Rough-set)</td>
<td>12.43</td>
<td>4.567</td>
</tr>
</tbody>
</table>

Table 3.2 portrays the performance analysis of the feature selection process using the Hadoop framework in the graphical portrayal. Here, Rough-set method and Random Forest tree are used into the Hadoop environment to discover the runtime of the Mapper phase and Reducer phase. As appeared in the table, obviously the time complexity nature lessening in the Mapper and Reducer stages of the Rough set with Random forest model contrasted with other classification techniques without a rough-set method.

![Hadoop Measures](image)

Figure 3.4 Algorithms with Map-Reduce Statistics

Figure 3.4 portrays the performance analysis of the feature selection process using the Hadoop framework in the graphical portrayal. Here, Rough-set method and Random Forest tree are used into the Hadoop environment to discover the runtime of the Mapper phase and Reducer phase. As appeared in the table, obviously the time complexity nature lessening in the Mapper and Reducer stages of the Rough set with
Random forest model contrasted with other classification techniques without a rough-set method.

Table 3.3 Classification models with True Positive and Error Rates metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>True Positive (%)</th>
<th>Error (%)</th>
<th>Outlier (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA_FSA</td>
<td>81.45</td>
<td>25.67</td>
<td>12.54</td>
</tr>
<tr>
<td>Neural Network</td>
<td>83.157</td>
<td>21.56</td>
<td>15.37</td>
</tr>
<tr>
<td>ELM-Tree</td>
<td>84.26</td>
<td>27.89</td>
<td>16.24</td>
</tr>
<tr>
<td>Random Forest(Roughset)</td>
<td><strong>89.45</strong></td>
<td><strong>19.04</strong></td>
<td><strong>11.15</strong></td>
</tr>
</tbody>
</table>

Table 3.3 and Figure 3.5 depict the order models' performance analysis. Random Forest tree with a roughset scores high genuine positive order rate for biomedical datasets when contrasted with other conventional classification models in Hadoop system.

![Figure 3.5 Performance metrics for classification models](image)

True positive rate demonstrates the quantity of positive instances reflects the attacks contrasted with negative examples. Error (%) demonstrates the quantity of
misclassified cases. Likewise, the outliers (%) which demonstrate the quantity of instances that are not pertinent to the current attacks conduct are additionally computed.

**Summary:**

In this chapter, as the size of distributed databases, data pre-processing and classification true positive rate are exponential decreased and they are different from the existing resolutions that necessitate an earlier learning of classification accuracy for a few sorts of information attributes, which is hard to accomplish in all reality. The fundamental issues of the conventional data mining methods over enormous data are distributed noise or anomaly problem, mining sparse issue and scaling up for high-dimensional space issue, and static optimization issue. In this chapter, a novel roughset based decision tree model using Hadoop framework was implemented on biomedical databases to improve the true positive rate and the computation error.