Cancer research is one of the major areas in the medical field. The accurate prediction of different tumour types has great value in providing better treatment and toxicity minimization of the patients. Until now, cancer classification has been based primarily on morphological symptoms and clinical based tumour. This has serious limitations because of ambiguity. These conventional cancer classification methods are reported to have several limitations in their diagnostic capability. It has been suggested that specifications of therapies according to the tumor types differentiated by pathogenetic patterns, may maximize the efficacy of the treatment of the patients. Also, the existing tumour classes have been found to be heterogeneous, and comprise of diseases that are molecularly distinct and follow divergent clinical courses.

Discovering the genes or features that are most relevant to the disease in question, and identifying the disease subtypes from such heterogeneous data remain an open problem. Due to the large variability in the gene mutations and gene expression in this population, till date not all patients have the same response to therapy, and pose a high challenge to physicians for treatment. The expression levels of genes are known to contain the keys to address fundamental problems relating to the prevention and treatment of diseases, biological evolution mechanisms and drug discovery.
In order to gain a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis have been designed. The present study is aimed to evaluate, using the microarray gene expression data of human acute leukemia, and to distinguish between ALL and AML, which is a typical cancer classification problem, not effectively solved despite research carried out over many years.

This research is aimed at the classification part of this problem using the two datasets of standard leukemia for training and testing, and the performance of the developed technique on clustering the ground truth data of the cancer classes, namely, AML and ALL is demonstrated. This high-dimensional training dataset is subjected to a multi-stage clustering technique, which performs clustering at diverse levels.

The findings of the study indicated that the developed technique is faster when compared to the existing clustering techniques in terms of performance. In addition, the method developed in this study also helps to reduce the data size, thus improving the running time. The experimental results based on three real datasets have demonstrated, that this developed technique is truly more robust and efficient, than traditional hierarchical clustering.

This study classified the AML and ALL types of cancer. AML is considered to be an aggressive cancer, and patients are often at a high risk of developing a cancer recurrence following therapy, particularly, if they are not able to undergo high doses of therapy. Chromosomal variables of AML cells, as well as levels of cancer cells in the blood, prior hematologic disorders and levels of specific enzymes, may help to further distinguish patients into being at a high-risk, standard-risk or low-risk of developing a cancer recurrence, and treatment may be altered according to these stratifications. This developed multi stage clustering approach helps in
determining the disease characteristics, and ultimately providing a platform to create individual treatment regimens. Further, through this method, one can identify different gene profiles that aid in predicting the risk of a cancer recurrence and/or the response potential to specific therapies. Moreover, the clustering approach also helps in predicting the survival probability of more than 57%.

7.1 CONTRIBUTIONS OF THIS RESEARCH

- A two-dimensional hierarchical, semi-supervised and quad tree clustering technique is developed and implemented, to handle the microarray gene that exists in more than one cluster.

- This method is used to evaluate the microarray gene expression data of ALL/AML, and the target is to distinguish between the ALL/AML. Thus, the cancer classification problem can be solved to a certain extent.

- The developed two dimensional hierarchical technique is also implemented with three real life datasets, namely, human acute leukemia, adenocarcinoma and lymphoma cancer cells.

- The inner gene clustering and hierarchical clustering are used in both the local and global information, and the local pairwise similarity measure is computed between the gene representations.

- This method performs inner gene clustering to discover overlapping clusters, which made biologically meaningful clusters.
• The most prominent gene expression data has been selected by this approach using the index matrix, which has reduced the biological complexities of working on the whole database.

• The minimum precision range of this developed approach is approximately closer to the maximum precision range of the FCM, FKM, SOM, Hybrid FCM, hierarchical clustering, kernel and GA based clustering algorithms.

• The minimum F-measure range of this developed method is approximately closer to the maximum F-measure range of the FCM, FKM, SOM, Hybrid FCM, hierarchical clustering, kernel and GA based clustering algorithms.

• The two-dimensional hierarchical clustering has achieved more precision, recall and F-score values over the existing clustering techniques.

• In comparison to the other hierarchical clustering methods, the performance of the quad tree based two dimensional clustering is higher, as it reduces the data size.

• In comparison to the other methods, the quad tree enhanced the interpretability, and visualization of the clustering results of the ALL/AML microarray gene expression data, which in turn, avoids biological complexities.

• The analytical results showed that the two dimensional hierarchical clustering technique using the hybrid similarity measure has better precision, recall and f-measure values of 8.710542%, 18.92406% and 17.93977% respectively, than those of prior research.
The semi-supervised two dimensional hierarchical clustering was developed using Matrix Laboratory, with the microarray gene expression data of human acute leukemia, adenocarcinoma and lymphoma cancer cells. However, with the same platform, the fast quad-tree based two dimensional hierarchical clustering and two dimensional hierarchical clustering with the hybrid similarity measure were implemented in human acute leukemia. The performance is compared with that of other unsupervised clustering algorithms, such as Hierarchical Clustering, Fuzzy C-mean clustering, Fuzzy K-mean clustering, Self-Organizing Maps, and Hybrid FCM and hierarchical clustering. The performances were evaluated through precision, recall, F-measure and Rand Index.

The huge number of genes and the intricacy of biological networks have highly increased the challenges of comprehending and interpreting the resulting group of data, increasing the processing time. The developed technique focuses on a QT based two dimensional hierarchical clustering algorithm to perform clustering. The construction of the closest pair data structure at each level is an important time factor, which determines the processing time of clustering. This developed model reduces the processing time, and improves the analysis of the gene expression data. The experimental results on real life datasets proved that the developed techniques are more effective than those of conventional hierarchical clustering.
7.2 FUTURE ENHANCEMENTS

In future gene expression analysis would help to classify patients according to the type of cancer, ultimately leading to a more individualized treatment approach, to improve the chances of optimal long-term survival.

Moreover, the advanced multi-stage clustering technique helps to improve the cancer disease classification, prognosis and prediction of response. The development of such methods in future will help to increase the interpretability, and visualization of the clustering results, as well as enhance the robustness and accuracy of the analysis by reducing the noise levels, and outliers, thereby providing better results in comparison with conventional clustering techniques.